

# A technique for constructing a DNA library encoding a structurally diverse repertoire of constrained peptides

Palmer, Simon James

*Award date:*
1998

*Awarding institution:*
University of Bath

# A technique for constructing a DNA library encoding a structurally diverse repertoire of constrained peptides

submitted by Simon James Palmer
for the degree of PhD
of the University of Bath
1998







ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

**$2^{3021377} - 1$ is prime**

# ABSTRACT

A technique for constructing a DNA library encoding a population of variable-length peptides ( ≤ 30 amino acids in length) which contain an even number of randomly distributed cysteine residues (2, 4, 6 or 8) amongst fixed amino acids, was successfully conceived and tested (Palmer *et al.*, 1998). Once constructed, this DNA library should encode a structurally diverse repertoire of ~$10^7$ constrained peptides (similar in many respects to the conotoxins), that would be a potentially rich source of novel peptide ligands and pharmaceutical leads.

The technique hinges upon the DNA ligase-catalysed polymerisation of six classes of small sticky-ended double-stranded DNA block. The crux of the technique is a three nucleotide 5′ overhang scheme which ensures that the assembled genes encode an even number of cysteines (thus the encoded peptides have the ability to form an integral number of disulphide bonds) and that the joint between two ligated blocks encodes a cysteine residue. Spacers in the dsDNA blocks encode a variable number of amino acids, allowing the length of the encoded peptide to vary and the random distribution of the cysteines. The technique also enables genes encoding peptides larger than 30 amino acids and the vast majority encoding more than 8 cysteines to be excluded from the DNA library. The technique works well although it does have a number of limitations which do not greatly affect its success overall.

## Acknowledgements

I would like to thank my supervisor Dr Jonathan Cox for his help and advice throughout this project. In addition, I would like to thank Rachel Sixsmith, Steve Searle and Phil Williams for their help with the illustrations in this thesis, David Palmer and Paul Plater for their help with the photographs, and my colleagues in laboratory 3.36/0.34.

This project was funded by the Department of Chemistry, University of Bath.

# TABLE OF CONTENTS

# Abbreviations

| | |
|---|---|
| $A_{260}$ | absorbance at 260nm |
| ATP | adenosine 5′-triphosphate |
| bp | base pairs |
| BSA | bovine serum albumin |
| $Ca^{2+}$-channel | calcium channel |
| CDR | complementarity determining region |
| $C_H$ | heavy chain, constant domain |
| $C_L$ | light chain, constant domain |
| dATP | deoxyadenosine 5′-triphosphate |
| dCTP | deoxycytidine 5′-triphosphate |
| dGTP | deoxyguanosine 5′-triphosphate |
| DNA | deoxyribonucleic acid |
| dNTP | deoxynucleoside 5′-triphosphate |
| dsDNA | double-stranded DNA |
| DTT | dithiothreitol |
| dTTP | deoxythymidine 5′-triphosphate |
| EDTA | disodium ethylenediamine tetraacetate |
| Fc | non-antigen-binding immunoglobulin fragment |
| IgG | immunoglobulin G |
| $K_a$ | association (or stability) constant |
| kb | kilobase pairs |
| $K^+$-channel | potassium channel |
| kDa | kilodaltons |
| LACI | lipoprotein-associated coagulation inhibitor |
| *lacZ* | *E. coli* β-galactosidase gene |
| LMP | low melting point |
| MCS | multiple cloning site |
| $Na^+$-channel | sodium channel |
| nAChR | nicotinic acetylcholine receptor |
| NEBuffer | New England Biolabs buffer |
| nt | nucleotides |

| | |
|---|---|
| PCR | polymerase chain reaction |
| PDB | Brookhaven protein data bank |
| PelB | peptide leader sequence from the *pelB* pectate lyase gene |
| rbs | ribosome binding site |
| ssDNA | single-stranded DNA |
| TAE | 40mM Tris-acetate, 1mM EDTA |
| *Taq* | thermostable DNA polymerase from *Thermus aquaticus* |
| TBE | 90mM Tris-borate, 2mM EDTA |
| TEMED | N, N, N′, N′-tetramethylethylene-diamine |
| Tris | tris-(hydroxymethyl)-methylamine |
| UV | ultra-violet |
| $Vent_R$ | thermostable DNA polymerase from *Thermococcus litoralis* |
| $V_H$ | heavy chain, variable domain |
| $V_L$ | light chain, variable domain |

# CHAPTER 1

# INTRODUCTION

This project concerns the construction of a DNA library encoding a structurally diverse repertoire of constrained peptides, which should be a rich source of novel peptide ligands and pharmaceutical leads. This introduction will therefore detail the basic paradigms of biological molecular recognition, the means by which the design principles of the conotoxins (the smallest units of biological molecular recognition) can be used to generate a structurally diverse repertoire of constrained peptides, and the efficacy of such a peptide repertoire. In addition, methods of expressing peptide repertoires encoded by DNA libraries and of selecting individual peptides from them are detailed.

## 1.1 NATURAL RECOGNITION MOLECULES

### 1.1.1 Human immunoglobulin G

**Structure and function**

Human immunoglobulin G (IgG) antibody molecules (150 kDa) consist of two identical heavy (H) chains (50 kDa) and two identical light (L) chains (25 kDa), which associate by disulphide and noncovalent bonding to produce a Y-shaped recognition molecule which is bivalent and mono-specific (Figure 1.1)(Kuby, 1997). The chains are composed of distinct domains: each light chain consists of a variable ($V_L$) and a constant domain ($C_L$); each heavy chain consists of a variable ($V_H$) and three constant domains ($C_H1$, $C_H2$ and $C_H3$). Each domain forms a characteristic tertiary structure known as an immunoglobulin fold, consisting of a beta-sheet "sandwich" containing either nine (V domains) or seven (C domains) anti-parallel beta-strands of amino acids.

Amino acid sequence variability within each of the $V_L$ and $V_H$ domains is concentrated in three hypervariable regions, which are located on the loops connecting the beta-strands, and which emanate from the domains (Figure 1.2). On association of the $V_L$ and $V_H$ domains the six hypervariable loops or complementarity-determining regions (CDRs) (which are not strictly equivalent but for this explanation they can be thought of as being so) are brought into proximity to constitute the antigen-binding site, and define the specificity of the antibody by producing a surface which is complementary in shape to an antigen. The surface area of contact between each antigen binding site and antigen is approximately 650-900 $Å^2$, within which approximately 15-22 amino acids contact a similar number or residues in the protein antigen. The main chain conformation of each CDR is determined by the loop length and a few key amino acid residues (Chothia and Lesk, 1987), whilst sequence variation in the CDR modulates the surface presented to the antigen. The remainder of the $V_L$ and $V_H$ domains are known as the framework regions and act as a rigid scaffold supporting the CDRs, whilst the constant domains are concerned with the biological effector functions of the immunoglobulin. Sequence variation in the CDR and framework regions can also shift the CDRs relative to one another by small but significant amounts (Chothia *et al.*, 1989).

**Mode of generation**

During human B-cell development, random somatic gene rearrangements involving a relatively small number of immunoglobulin genes (Tonegawa, 1983) generate a diverse primary antibody repertoire of greater than $10^8$ different antigenic specificities. The size and diversity of this naïve repertoire (i.e. the repertoire before antigen challenge) ensures that any foreign antigen can be recognised by antibodies of the immune system, albeit with relatively low affinity, $K_a \sim 10^6$-$10^7$ $M^{-1}$ (Clackson and Wells, 1994). Note that each mature B-cell expresses surface bound immunoglobulin (the IgM and IgD classes) of a single antigenic specificity.

Antigenic stimulation of a mature B-cell induces cell proliferation and somatic hypermutation, a process by which the specificity of the expressed antibody is

apparently randomly altered by hypermutation of the genes encoding the variable domains, especially in the regions encoding the CDRs (Berek and Milstein, 1987)(Wagner *et al.*, 1995). Although the mechanism of hypermutation is unknown, an error-prone DNA polymerase has been suggested (Brenner and Milstein, 1966). The resulting population of antibodies with varying affinity for antigen is referred to as the secondary antibody repertoire. Antigen-mediated selection of high affinity antibodies from the secondary antibody repertoire, in a process akin to evolution known as affinity maturation (Rajewsky, 1996), ensures that the antigen-specific antibody from the naïve repertoire is matured to high affinity, $K_a \geq 10^9\ M^{-1}$ (Clackson and Wells, 1994). Additional rearrangement of the heavy-chain constant genes, in a process known as class switching, ensures the expression and secretion of the IgG class of antibody with the same antigenic specificity.

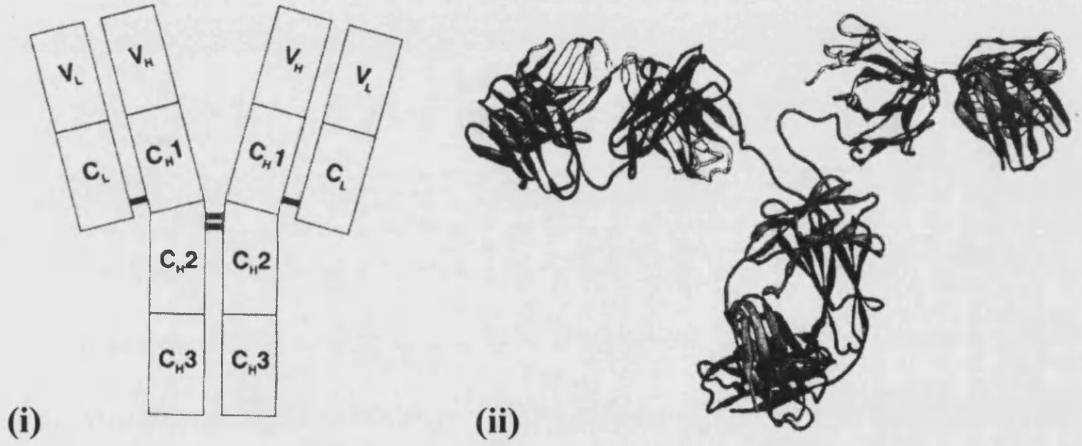



**Figure 1.1** Cartoon (i) and ribbon (ii) structural representation of human IgG. The heavy and light chains associate to produce a bivalent and mono-specific recognition molecule. Each light chain consists of a variable ($V_L$) and constant domain ($C_L$), whilst each heavy chain consists of a variable ($V_H$) and three constant domains ($C_H1$, $C_H2$ and $C_H3$). The antigen-binding sites are formed after association of the $V_L$ and $V_H$ domains brings the six CDRs into proximity. Figures reproduced from (i) Rees *et al.* (1994) and (ii) Borrebaeck (1995).

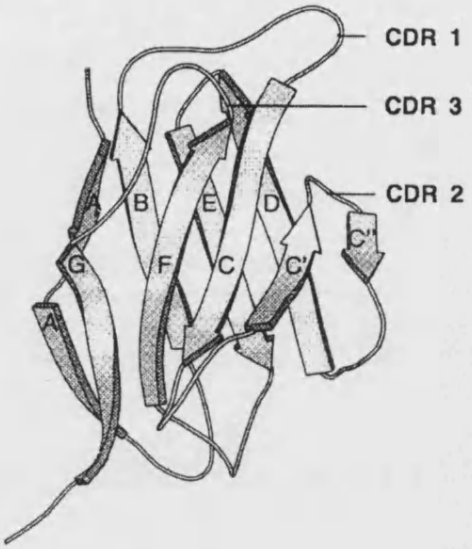



**Figure 1.2** Ribbon structural representation of an IgG variable domain. The nine beta-strands of a variable domain fold to form a beta-sheet sandwich. The loops connecting the nine beta-strands form the CDRs, which emanate from the domains and are denoted here arbitrarily as CDR1, CDR2 and CDR3. On association of the $V_L$ and $V_H$ domains the six CDRs are brought into proximity to constitute the antigen-binding site, and define the specificity of the antibody. However, it is interesting to note that natural antibody variable domains presenting only three CDR loops can successfully bind antigen (e.g. in camel antibodies which lack light chains)(Desmyter *et al.*, 1996). Figure reproduced from Rees *et al.* (1994).

### 1.1.2 Conotoxins

The cone snails (*Conus*) are a large genus of venomous marine predators that inject their prey with a venom that causes rapid paralysis, enabling the cone snails to capture and consume fish and other organisms (Olivera *et al.*, 1990a, 1990b, 1991, 1995). The *Conus* venom causes paralysis by selectively binding to a diverse set of ion channels and receptors critical to the functioning of the prey's neuromuscular system. The biologically active components of the *Conus* venoms are small, cysteine-rich peptides termed conotoxins that are between 7-34 amino acids in length (approximately 1-4 kDa). The conotoxins are conformationally constrained by 1, 2 or 3 disulphide bonds, enabling them to adopt a rigid and stable three-dimensional structure. The minimum size of a stable protein domain lacking disulphide bonds is approximately 50 amino acids (Privalov and Gill, 1988). It is interesting to note that

some conotoxins with no disulphide bonds at all, the conantokins, are able to adopt a defined and stable conformation in the presence of calcium ions, for example conantokin G adopts a distorted $3_{10}$ helix (Rigby *et al.*, 1997).

Since there are over 500 *Conus* species, and each one has a characteristic complement of up to 100 unique conotoxins within its venom, a multiplicity of conotoxins exist in nature, of which the amino acid sequences of approximately 65 have been published. Despite the natural diversity, a striking feature of the conotoxins is that they utilise one of a small number of highly conserved disulphide frameworks (Figure 1.3), although the disulphide frameworks utilised by the "four-loop" δ- ω- κ- μO-conotoxins [1], the "three-loop" μ- ψ-conotoxins [2] and the "two-loop" α-conotoxins [3] represent the three principal ones. Consequently, conotoxins are classified on the basis of their biological activity and their disulphide framework.

The folded main chain conformation of a conotoxin is determined by its length and the distribution of its disulphide-bonded cysteine residues; the remaining amino acid composition seems to have little influence (Hu *et al.*, 1996)(Lancelin *et al.*, 1991)(Mitchell *et al.*, 1996)(Farr-Jones *et al.*, 1995)(Kohno *et al.*, 1995)(Pallaghy *et al.*, 1993)(Mitchell *et al.*, 1998). The exception to this statement is that a glycine residue is strongly conserved at a certain position in the δ- and ω-conotoxins (Olivera *et al.*, 1990b, 1995) where it plays an important role in defining their structure (Pallaghy *et al.*, 1993). The disulphide framework of a conotoxin forms a structural core from which the inter-cysteine loops project (Figure 1.4). The emanating loops define the specificity of the conotoxin in a manner reminiscent of the CDRs of antibodies (Section 1.1.1), with a few key residues determining the biological activity (Sato *et al.*, 1991)(Hill *et al.*, 1996).

The strategy of *Conus* is to vary the length and sequence of the inter-cysteine loops, thereby altering the specificity of the conotoxin, whilst rigidly conserving the three-dimensional structure defined by the disulphide framework. The inter-cysteine loop length generally ranges from 0-7 amino acids, although loops of 9 amino acids are found in some conotoxins. The mechanism by which *Conus* generate hypervariability in the inter-cysteine loops of the disulphide frameworks is unknown, although random

mutation of the conotoxin genes followed by selection, or a cassette switching mechanism have been suggested (Woodward *et al.*, 1990). Conotoxins are therefore analogous to antibodies in the sense that they have constant regions defining their structure, and hypervariable regions defining their specificity (Section 1.1.1).

Conotoxins achieve the highest density of disulphide-bonding thus encountered in any biological system (Woodward *et al.*, 1990), and represent the smallest DNA-encoded peptide structures in nature functioning as recognition molecules (Olivera *et al.*, 1995). In comparison, peptide toxins from other venomous animals such as snakes, scorpions and spiders are much larger than the conotoxins, being approximately 35-75 amino acids in size (Swanson, 1996). Conotoxins demonstrate that constrained peptides have the ability to form highly specific high-affinity molecular interactions. For example, ω-conotoxin GVIA (Figure 1.3) binds to mammalian neuronal $Ca^{2+}$-channels with an exceptionally high affinity, $K_a = 1.2 \times 10^{12}\ M^{-1}$ (Barhanin *et al.*, 1988). It has been hypothesized that the cone snails only need to use a few conotoxin disulphide frameworks to fulfill their biological requirements (Woodward *et al.*, 1990) as the targeted ion channels and receptors belong to structurally related families (Smith, 1996).

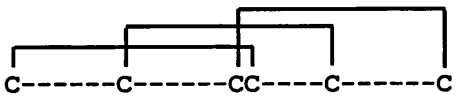
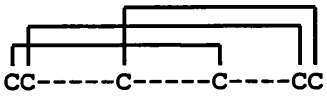
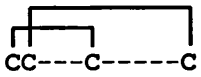
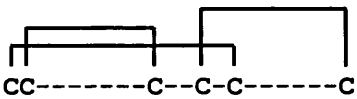
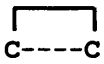
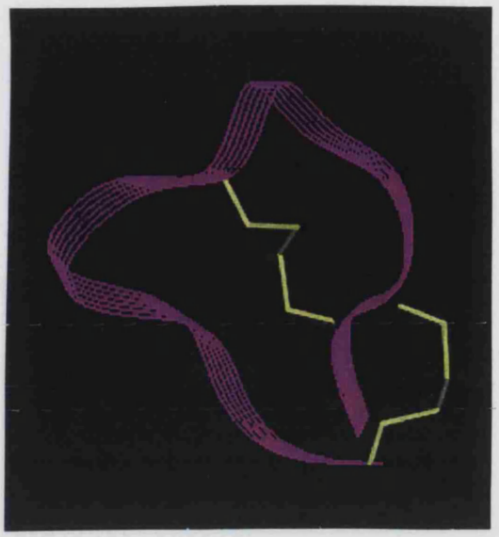
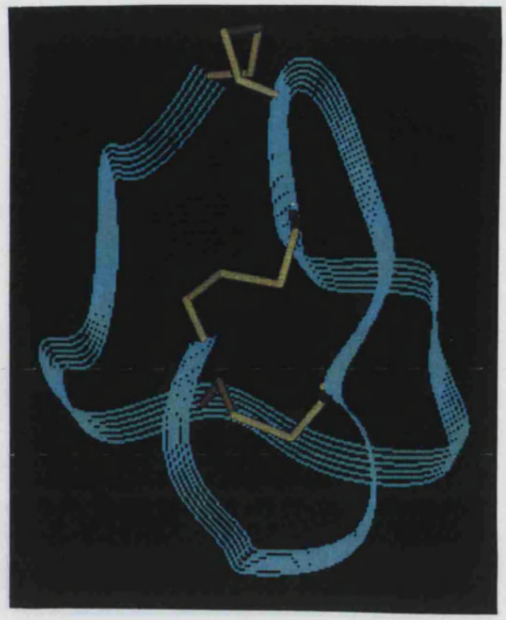
| Conotoxin | Sequence | Target | Species |
|---|---|---|---|
| **[1]** |  | | |
| | C------C------CC----C------C | | |
| δ-TxVIA | WCKQSGEMCNLLDQNCCDGY-CIVLV--CT | $Na^+$-channel | *C. textile* |
| δ-GmVIA | VKPCRKEGQLCDPIFQNCCRGWNCVLF---CV | | *C. gloriamaris* |
| ω-GVIA | CKSOGSSCSOTSYNCCRS--CNOYTKRCY* | $Ca^{2+}$-channel | *C. geographus* |
| ω-MVIIC[1] | CKGKGAPCRKTMYDCCSGS-CGRRGK-C* | | *C. magnus* |
| κ-PVIIA[2] | CRIONQKCFQHLDDCCSRK-CNRFNK-CV | $K^+$-channel | *C. purpurascens* |
| μO-MrVIB[3] | ACSKKWEYCIVPILGFVYCCPGLICGPFV--CV | $Na^+$-channel | *C. marmoreus* |
| **[2]** |  | | |
| | CC-----C-----C----CC | | |
| μ-GIIIA | RDCCTOOKKCKDRQ-CKOQRCCA* | $Na^+$-channel | *C. geographus* |
| μ-GIIIB[4] | RDCCTOORKCKDRR-CKOMKCCA* | | *C. geographus* |
| ψ-PIIIE[5] | HOOCCLYGK-CRRYOGCSSASCCQR* | nAChR | *C. purpurascens* |
| **[3]** |  | | |
| | CC---C-----C | | |
| α-GI | ECCNPACGRHYSC* | nAChR | *C. geographus* |
| α-SIA | YCCHPACGKNFDC* | | *C. striatus* |
| **[4]** |  | | |
| | CC-------C--C-C------C | | |
| αA-PIVA[6] | GCCGSYONAACHOCSCKDROSYCGQ* | nAChR | *C. purpurascens* |
| **[5]** |  | | |
| | C----C | | |
| Conopressin G[7] | CFIRNCPKG* | Unknown | *C. geographus* |
| Conopressin S[7] | CIIRNCPRG* | | *C. striatus* |
| **[6]** | CC-----C---C--C-C | | |
| μPnIVA[8] | CCKYGWTCLLGCSPCGC | $Na^+$-channel | *C. pennaceus* |
| μPnIVB[8] | CCKYGWTCWLGCSPCGC | | *C. pennaceus* |

**Figure 1.3** Six conotoxin disulphide frameworks. The connectivity of each disulphide framework is graphically illustrated, the cysteine arrangement of each conotoxin example is highlighted along with the species from which it is obtained, and the pharmacological target of each conotoxin class is detailed. Hyphens indicate gaps in the sequence alignment whilst an underlined conotoxin denotes that its structure is detailed in Figure 1.4. Conotoxin μO-MrVIB is not perfectly aligned as it contains an inter-cysteine distance of 9 amino acids. Note that O represents 4-*trans*-hydroxyproline and * denotes peptides that are amidated at the C terminus (a definitive explanation of the role of post-translational modifications in conotoxins remains unknown [Hill *et al.*, 1997]). The disulphide connectivity for framework [6] has not been determined (thus it may not be novel) and the μO- and δ-conotoxins are classified separately because they exhibit different pharmacological activity towards the $Na^+$-channel. Table reproduced with alterations from Olivera *et al.* (1995). Superscript number indicates reference: [1](Farr-Jones *et al.*, 1995), [2](Shon *et al.*, 1998), [3](McIntosh *et al.*, 1995), [4](Hill *et al.*, 1996), [5](Mitchell *et al.*, 1998), [6](Han *et al.*, 1997), [7](Gray *et al.*, 1988) and [8](Fainzilber *et al.*, 1995).

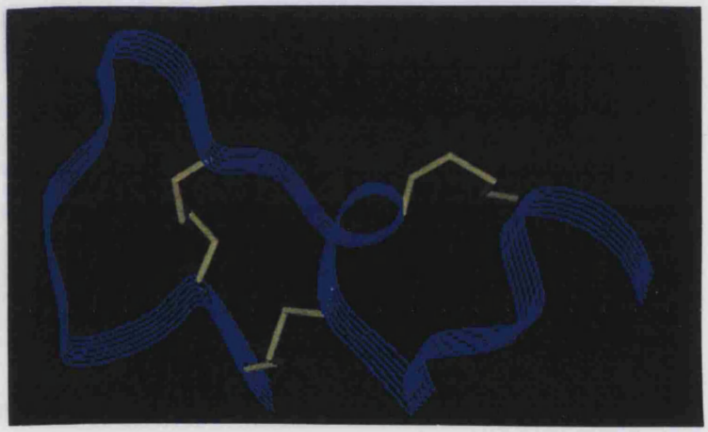


**i)** α-conotoxin GI (E<u>CC</u>NPA<u>C</u>GRHYS<u>C</u>*) from the cone snail *C. geographus* (PDB code: 1NOT; Guddat *et al.*, 1996) which targets nAChR and utilises a "two-loop" disulphide framework.

**ii)** μ-conotoxin GIIIB (RD<u>CC</u>TOORK<u>C</u>KDRR<u>C</u>KOMK<u>CC</u>A*) from the cone snail *C. geographus* (PDB code: 1GIB; Hill *et al.*, 1996) which targets sodium channels and utilises a "three-loop" disulphide framework.



**iii)** ω-conotoxin MVIIC (CKGKGAPCRKTMYDCCSGSCGRRGKC*) from the cone snail *C. magnus* (PDB code: 1OMN; Farr-Jones *et al.*, 1995) which targets calcium channels and utilises a "four-loop" disulphide framework.



**iv)** αA-conotoxin PIVA [Pro 7, 13] (GCCGSYPNAACHPCSCKDROSYCGQ*) from the cone snail *C. purpurascens* (PDB code: 1P1P; Han *et al.*, 1997) which targets nAChR.

**Figure 1.4** Ribbon structural representation of four conotoxins representing four disulphide frameworks. In each case the disulphide framework forms a structural core from which the inter-cysteine loops project to define the specificity of the conotoxin. The cysteine residues in the conotoxin sequences are underlined and the disulphide connectivity of each conotoxin is detailed in Figure 1.3. Note that [Pro 7, 13] αA-conotoxin PIVA is an equipotent analogue of αA-conotoxin PIVA with proline substitutions at positions 7 and 13, and is present as a major component of *C. purpurascens* venom (Hopkins *et al.*, 1995). The disulphide bonds are coloured yellow.

## 1.2 COMBINATORIAL PEPTIDE LIBRARIES

### 1.2.1 Combinatorial chemistry

Combinatorial chemistry involves the synthesis of a large and diverse repertoire of molecular variants, followed by screening or selection of the variants on the basis of desired activity towards a chosen target (Hogan, 1996)(Verdine, 1996). The premise of combinatorial chemistry is that as the size and diversity of the repertoire increases, so does the probability of finding a variant with the desired activity. Crucially, and in contrast to structure-based rational design, no detailed knowledge of the target molecule is necessary.

### 1.2.2 Construction and selection

The ability to construct large and diverse repertoires of peptide variants on a basic molecular theme ("combinatorial peptide libraries"), coupled with powerful means of selecting individual peptides from them, enable artificial peptide or protein "ligands" to be generated with *de novo* binding activities towards "targets" (Fisch *et al.*, 1996). Note that the most common type of combinatorial peptide library is one in which a number of residues in a peptide sequence have been randomised.

The construction of a biological combinatorial peptide library involves the synthesis of a large and diverse repertoire of genes, followed by expression of the encoded peptides. The most common expression format is display on filamentous bacteriophage ("phage display") by fusion to the N-terminus of the phage coat proteins pIII or pVIII, which are present at 3-5 and ~2700 copies per phage, respectively (Smith and Petrenko, 1997)(Smith, 1985). Selection of the peptide library on the basis of affinity towards a chosen target is achieved by affinity purification of the phage library on an immobilised target. Since the gene encoding the fusion protein is part of the phage genome, once a phage expressing a target-specific peptide has been isolated, the DNA sequence encoding the peptide can be determined and the amino acid sequence of the peptide can be deduced. Phage

display therefore provides direct linkage of genetic information with encoded selectable function (Noren, 1996). It is important to note that the maximum size of a conventional phage library is limited to $10^8$ by the transformation efficiency of *E. coli* (Waterhouse *et al.*, 1993).

As peptide-geneVIII libraries display 100-200 copies of fusion protein per phage, multivalent attachment of the phage to the immobilised target protein facilitates the capture and isolation of weakly-binding (i.e. low-affinity) peptides (e.g. $K_a = 10^4$–$10^6$ $M^{-1}$ [Clackson and Wells, 1994])(Wells, 1996). In contrast, as peptide-geneIII libraries display 0-5 copies of fusion protein per phage, monovalent attachment of the phage to the immobilised target protein facilitates the capture and isolation of strongly-binding (i.e. high-affinity) peptides (Wells, 1996)(Lowman *et al.*, 1991).

Fusion of displayed peptides to the N-terminus of the pIII or pVIII coat proteins via a flexible linker (e.g. GAGAA [Doorbar and Winter, 1994]), coupled with the fact that the N-terminus tends to be one of the most unconstrained parts of a protein (Richardson, 1981), should ensure that each displayed peptide adopts an autonomous fold. Consequently, peptides isolated by phage display should retain their activity towards the target molecule after detachment from the phage particle (Koivunen *et al.*, 1995)(Cunningham *et al.*, 1994)(Doorbar and Winter, 1994). The distance between neighbouring coat proteins on the phage surface should also prevent intermolecular disulphide-bond formation between displayed peptides (Smith and Petrenko, 1997).

### 1.2.3 Constrained peptides

Although large numbers of unconstrained random peptide libraries have been constructed and displayed on phage (Smith and Petrenko, 1997), there are many targets for which no high-affinity binding peptides can be found (Ladner, 1995)(Clackson and Wells, 1994).

Imposing conformational constraints upon peptides confers numerous advantages over flexible, unconstrained peptides (Ladner, 1995). A constrained peptide that presents a surface complementary in "shape" to its target should have a higher binding

affinity than an unconstrained peptide, as both factors favourably influence the free energy change accompanying complex formation. Structural complementarity encourages enthalpic contributions from hydrogen bonds, ionic bonds, hydrophobic interactions and van der Waals interactions (Kuby, 1997), whilst rigidity minimises adverse contributions from conformational entropy losses. Other advantages of constrained peptides are that they can present residues in an unfavourable environment (e.g. hydrophobic residues), and that the difference in affinity between a phage-bound and phage-free peptide is relatively small. Thus the tighter a peptide is conformationally constrained, the less likely it is to bind to a chosen target (since its conformation is fixed), although if binding does occur it is likely to be highly specific and to have a high affinity. Most peptides that bind specifically to target proteins *in vivo* are constrained, usually by disulphide-bonding (Clackson and Wells, 1994).

The most common conformational constraint imposed on displayed peptides is a single disulphide bond between two cysteine residues at fixed positions in an otherwise random sequence [$CX_ZC$ : where C is cysteine, X is a randomised residue and Z represents the number of randomised residues], giving a library of random cyclic peptides (Smith and Petrenko, 1997). In one study peptides isolated from libraries of random cyclic peptides [$CX_4C$, $CX_5C$ and $CX_6C$] bound streptavidin with affinities up to 3 orders of magnitude higher than peptides isolated from unconstrained peptide libraries (Giebel *et al.*, 1995). In another study a library of random cyclic peptides [$CX_7C$] was screened against several integrins and the strongest binder [CDCRGDCFC] actually contained a second disulphide bond, which presumably constrained the peptide even further (Koivunen *et al.*, 1995).

Unconstrained linear peptides and constrained cyclic peptides have "continuous" binding sites, comprising a short continuous stretch of the peptide sequence. For example, antibodies specific for continuous binding sites on protein antigens typically contact 3-4 amino acids over a six residue segment (Smith and Petrenko, 1997). However, most molecular recognition events between peptides or proteins and their targets are mediated by "discontinuous" binding sites, comprising critical residues that are distant in the primary sequence but close in the folded conformation (Jones and Thornton, 1996). The classic example is the antigen-binding site of an immunoglobulin molecule, which is formed from six CDRs (Section 1.1.1).

### 1.2.4 Engineering of novel binding sites into protein scaffolds

A second method of constraining peptides is to present them in the context of a protein scaffold (Nygren and Uhlén, 1997), a scaffold being defined as a rigid structural support. Randomised surface residues can either be presented continuously (e.g. as a single loop) or discontinuously (e.g. as distributed over a suitable region of the molecular surface, such as the binding site of the native protein). Any protein displayed on phage is therefore a potential scaffold for the presentation of a combinatorial peptide library. Combinatorial peptide libraries have been presented from protein scaffolds with the aim of improving the binding affinity of the protein with its natural receptor. For example, in one study the binding affinity of human growth hormone for its receptor was increased 400-fold to $K_a = 10^{12}$ $M^{-1}$ (Lowman and Wells, 1993). However, combinatorial peptide libraries presented from protein scaffolds have also been used to endow protein scaffolds with novel binding activity (Figure 1.5).

The advantages of presenting combinatorial peptide libraries from protein scaffolds are that the defined and stable structure of the scaffold conformationally constrains the presented combinatorial peptide library, and novel discontinuous binding sites can be created (Section 1.2.3). However, there are also a number of disadvantages. Firstly, the number of residues in a protein scaffold that can tolerate randomisation without disrupting the three-dimensional structure may be limited, and these residues may only comprise one face of the protein. Secondly, the three-dimensional structure of the native protein scaffold must be known so that the positions of these residues can be determined.

| Protein Scaffold | Size (Amino Acids) | Native Ligand or Activity | Number of Randomised Positions | Novel Specificity | Reference |
|---|---|---|---|---|---|
| **Minibody** (truncated $V_H$ domain) | 61 | Phosphorylcholine | 6 + 6 | Human interleukin-6 | (Tramontano *et al.*, 1994) (Martin *et al.*, 1994, 1996) |
| **CP-1 zinc finger** | 26 | Specific DNA | 5 | *S. flexneri* specific monoclonal antibody | (Choo and Klug, 1995) (Bianchi *et al.*, 1995) |
| **Tendamistat** | 74 | α-amylase inhibitor | 3 + 6 | Endothelin specific monoclonal antibody | (McConnell and Hoess, 1995) |
| **Z domain of protein A** | 58 | Immunoglobulin G (Fc region) | 13 | Human insulin, Taq DNA polymerase, Apolipoprotein A-1 | (Nord *et al.*, 1995, 1997) |
| **Pancreatic secretory trypsin inhibitor** | 56 | Trypsin inhibitor | 7 or 8 | Chymotrypsin | (Röttgen and Collins, 1995) |
| **Kunitz inhibitor domain of human LACI** | 58 | LACI | 5 + 4 | Plasma kallikrein, Thrombin, Plasmin | (Markland *et al.*, 1996a, 1996b) |
| **Cytochrome $b_{562}$** | 106 | Electron transport protein | 5 + 4 | N-methyl-p-nitrobenzylamine-BSA conjugate | (Ku and Schultz, 1995) |

**Figure 1.5** Protein scaffolds engineered with novel binding sites. Where more than one region in the protein scaffold has been randomised this is denoted by a + sign along with the number of randomised residues in each region. Reproduced from Nygren and Uhlén (1997).

Note that libraries of antibody fragments in which the CDRs have been randomised and in which different CDRs have been put together in a combinatorial manner, have successfully generated antibodies with novel binding specificities (Griffiths *et al.*, 1994)(Winter *et al.*, 1994).

### 1.2.5 Engineering of novel binding sites into peptide scaffolds

A number of features make constrained peptides attractive scaffolds for the presentation of combinatorial peptide libraries, which can endow them with novel binding activity.

Firstly, as the three-dimensional structure of a constrained peptide is defined by the arrangement of its disulphide-bonded cysteine residues (Section 1.1.2), the vast majority (if not all) of the non-cysteine residues in a constrained peptide can tolerate randomisation without disrupting the overall structure. Consequently, the structure of a constrained peptide does not have to be determined for it to be used as a scaffold, and constrained peptide scaffolds can be much smaller in size than their protein counterparts (Section 1.2.4).

Secondly, as the non-cysteine residues project from a constrained peptide as loops (Figure 1.4), the whole molecular surface of a constrained peptide can be randomised, enabling continuous and discontinuous novel binding sites to be created, and different faces of the constrained peptide to interact with different targets (Smith *et al.*, 1998). For example, four inter-cysteine loops emanate from the μO-, δ-, κ- and ω-conotoxins in three dimensions (Figure 1.4iii). Note that only a few randomised loops presented from a molecular scaffold are needed for successful molecular recognition (Figures 1.2 and 1.4).

Thirdly, a constrained peptide generally adopts a rigid and stable conformation that conformationally constrains the presented combinatorial peptide library (Section 1.2.3). However, note that the inter-cysteine loops of a constrained peptide (which present the combinatorial peptide library) will have approximately the same spatial relationship to the constrained peptide scaffold regardless of their amino acid composition (Ladner, 1995). It is important to retain the conformation of any peptide (or protein) scaffold in a combinatorial peptide library, so that the nature of any novel binding activity and the spatial relationship between the pharmacophoric groups can be easily determined; this aids the formation of small-molecule mimetics (see below) (Bianchi *et al.*, 1995)(Barbato *et al.*, 1996).

Finally, as molecular recognition is generally focused to a relatively small number of residues for a constrained peptide (Ladner, 1995), and since a peptide can be chemically synthesized and its structure determined relatively easily (Vita *et al.*, 1995), constrained peptides make attractive pharmaceutical leads for the design of

small-molecule mimetics (Section 6.3). Note that a pharmaceutical lead is defined as a molecule which affects a chosen target in the desired manner and which can possibly be developed into a pharmaceutical (Wells, 1996)(Broach and Thorner, 1996).

Natural constrained peptides have been successfully endowed with novel binding activity. In one work the 36 amino acid C-terminal cellulose binding domain (CBD) of cellobiohydrolase I, which is constrained by 2 disulphide-bonds, was endowed with novel binding activity towards alkaline phosphatase (Smith *et al.*, 1998). Seven residues in two regions, comprising the cellulose-binding site of the native peptide, were randomised in the combinatorial peptide library. In another work a 37 amino acid scorpion toxin (charybdotoxin), which is constrained by 3 disulphide-bonds, was endowed with a metal binding site after nine discontinuous residues of the native peptide were mutated (Vita *et al.*, 1995).

Theoretically, any natural constrained peptide could be used as a scaffold for the presentation of a combinatorial library to create peptides with novel binding activity. Ideal candidates include members of the cysteine-knot (Krätzner and Kolmar, 1997)(Pallaghy *et al.*, 1994)(Isaacs, 1995)(Harrison and Sternberg, 1996) and knottin (Smith *et al.*, 1998)(Le Nguyen *et al.*, 1990)(Lin and Nussinov, 1995) families of proteins, defensins (Harder *et al.*, 1997), conotoxins (Section 1.1.2) and other toxins (Swanson, 1996). The presentation of a combinatorial peptide library from a constrained peptide scaffold is effectively the strategy used by the cone snails to generate novel conotoxins (Section 1.1.2) and a combinatorial peptide library presented from an α-conotoxin scaffold has been successfully displayed on bacteriophage (Bonnycastle *et al.*, 1996).

However, the limitation of presenting a combinatorial peptide library from a constrained peptide representing a single constrained topology is that the peptide variants may only be compatible with certain target types (Clackson and Wells, 1994). In support of this point it may be noted that the ion channels and receptors targeted by each of the conotoxin disulphide-frameworks belong to structurally related families (Section 1.1.2). Also, the C-terminal cellulose binding domain (CBD) of

cellobiohydrolase I could not be endowed with novel binding activity towards α-amylase and β-glucuronidase (Smith *et al.*, 1998). Consequently, a better strategy may be to initially find a constrained peptide with a suitable constrained topology for the chosen target, and to subsequently present a combinatorial peptide library from that constrained peptide scaffold in order to optimise its binding affinity.

## 1.3 THE RATIONALE OF THIS PROJECT

### 1.3.1 A repertoire of constrained peptides

Highly specific, high affinity ($K_a \geq 10^9$ $M^{-1}$) molecular interactions between peptides and their targets occur when the peptide has a rigid conformation and presents a surface that is complementary in "shape" to the target (Section 1.1.2). As the "shape" of a peptide is dependent on its folded main chain conformation, and since constrained peptides generally adopt rigid and stable conformations, it was reasoned that a structurally diverse repertoire of constrained peptides would have great potential as a source of novel peptide ligands.

As the folded main chain conformation of a constrained peptide is determined by its length and the distribution of its disulphide-bonded cysteine residues (Section 1.1.2), variation of length and cysteine arrangement in a combinatorial peptide library should generate a structurally diverse repertoire of novel constrained peptides. The aim of this project is to determine a means of constructing a DNA library that encodes such a combinatorial peptide library (Section 1.4).

A number of physical limitations are to be imposed upon the combinatorial peptide library. Firstly, the peptides are to have a maximum size of 30 amino acids, in keeping with the conotoxins (Section 1.1.2), and as small peptides make good pharmaceutical leads (Section 1.2.5). Secondly, the peptides are to contain an even number of cysteine residues so that they have the ability form an integral number of disulphide bonds. Finally, the maximum number of cysteine residues in any peptide is to be limited to eight, in keeping with natural peptides of this size, to prevent the disulphide-bonding from becoming too dense.

It is important to note that the length of each peptide in the combinatorial peptide library is defined as the number of amino acids between the outermost cysteines inclusively. Therefore, according to this definition a peptide of 30 amino acids (the chosen maximum) contains a cysteine as its first and last residue. Note that conotoxins have relatively few amino acids beyond their outermost cysteines (Section 1.1.2). Although the maximum number of cysteines in the conotoxins is limited to six (Section 1.1.2), other natural peptides contain eight cysteines of which the outermost are separated by approximately 30 amino acids (Narasimhan *et al.*, 1994)(Hill *et al.*, 1997)(Olamendiportugal *et al.*, 1996)(Lebrun *et al.*, 1997).

Ideally, the repertoire of constrained peptides will be so diverse that for any given target a peptide of some affinity can be isolated. Such a peptide library would be analogous to the naïve primary antibody repertoire of the human immune system (Section 1.1.1) and would be a potentially rich source of pharmaceutical leads (Section 1.2.5).

This repertoire of constrained peptides, representing a multiplicity of constrained topologies, would effectively be a "library of libraries" from which a constrained peptide with a suitable constrained topology could be selected for any chosen target. Following selection, a second combinatorial peptide library could then be presented from this constrained peptide scaffold in order to improve its binding affinity towards the target (Section 6.3). Note that constrained peptides make excellent scaffolds for the presentation of combinatorial peptide libraries (Section 1.2.5).

## 1.4 THE AIM OF THIS PROJECT

The aim of this project is to determine a means of constructing a DNA library encoding a population of variable-length peptides ( $\leq 30$ amino acids in length) which contain an even number of randomly distributed cysteine residues (2, 4, 6 or 8). Such a DNA library should encode a structurally diverse repertoire of constrained peptides, that could be a rich source of novel peptide ligands and pharmaceutical leads.

# CHAPTER 2

# GENERAL METHODS

This chapter details the general experimental techniques used throughout this project. Any methods specific to a particular chapter are described within that chapter.

## 2.1 EXPERIMENTAL METHODS

### 2.1.1 Molecular biological materials

All restriction endonucleases, DNA polymerase I (large fragment [Klenow]), $Vent_R$ DNA polymerase, T4 DNA ligase, T4 polynucleotide kinase, β-agarase I and bovine serum albumin were supplied, along with the relevant buffers, by New England Biolabs, Beverly, USA. *Taq* DNA polymerase was obtained as BIOTAQ™ polymerase from Bioline Ltd, London, UK. The *E. coli* culture media constituents were supplied by Difco Laboratories, Michigan, USA. Ultrapure deoxynucleosidetriphosphates (dNTPs) were obtained from Pharmacia Biotech, Uppsala, Sweden. The antibiotics kanamycin and ampicillin were obtained from Sigma, UK. The Eukaryotic TA cloning kit (version 1.0) and the S.N.A.P.™ miniprep kit were obtained from Invitrogen BV, the Netherlands. The 50 bp and 1 kb DNA ladders were obtained from Life Technologies, Paisley, UK. A premade 37.5:1 acrylamide:bisacrylamide solution (called Protogel™), supplied at a 30% (w/v) acrylamide concentration, was obtained from Flowgen Instruments Ltd., Staffordshire, UK. Microcon™-10 microconcentrators and Micropure™-0.22μm inserts were obtained from Amicon Inc, Beverly, USA.

### 2.1.2 Bacterial strains and culture conditions

*Escherichia coli* strain TOP10F' (F' {*lac*I$^q$ *Tn*10(Tet$^R$)} *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L *end*A1 *nup*G) was used throughout this project, and was obtained as part of the Eukaryotic TA cloning kit.

*E. coli* strain TG1 (*sup*E hsdΔ5 *thi* Δ(*lac-proAB*) F'[*traD*36 *proAB*$^+$ *lacI*$^q$ *lacZ*ΔM15]) (Gibson, 1984)(Sambrook *et al.*, 1989) was the strain in which the pUC119- *Sfi / Not* - $HIS_6$ plasmid was supplied in (Section 2.1.4).

*E. coli* strains were cultured at 37°C in a rotary shaker (200 cycles/minute) in either 2xYT (16 g/l bacto-tryptone, 10 g/l bacto-yeast extract, 5 g/l NaCl) or LB (10g/l bacto-tryptone, 5g/l bacto-yeast extract, 10g/l NaCl) liquid media. *E. coli* strains were also cultured at 37°C on either 2xYT-agar or LB-agar solid medium (each supplemented with 15 g/l bacto-agar).

For the propagation of *E. coli* strains carrying the ampicillin and kanamycin resistance-conferring plasmid pCR3 (Section 2.1.3) the media was supplemented with ampicillin or kanamycin to a concentration of 50μg/ml and 25μg/ml, respectively. For the propagation of *E. coli* strains carrying the ampicillin resistance-conferring plasmid pUC119- *Sfi / Not* -$HIS_6$ (Section 2.1.4) the media was supplemented with ampicillin to a concentration of 100μg/ml.

Long term storage of *E. coli* was achieved by incubating cells from an overnight culture in 16% (v/v) glycerol at -70°C.

### 2.1.3 The pCR3 plasmid

The 5.1kb pCR3™ plasmid (Figure 2.1) used throughout this work was obtained as part of the Eukaryotic TA cloning kit. The pCR3 plasmid had been designed to

bidirectionally clone *Taq* DNA polymerase PCR products and to express them in mammalian cell lines under control of the cytomegalovirus immediate-early promoter (Akrigg *et al.*, 1985)(Boshart *et al.*, 1985). As *Taq* DNA polymerase has a nontemplate-dependent activity which adds single deoxyadenosine residues to the 3′ ends of duplex DNA molecules (Clark, 1988) the pCR3 plasmid was supplied as a 5058 bp linear vector with single 3′ deoxythymidine overhangs (Figure 2.2).

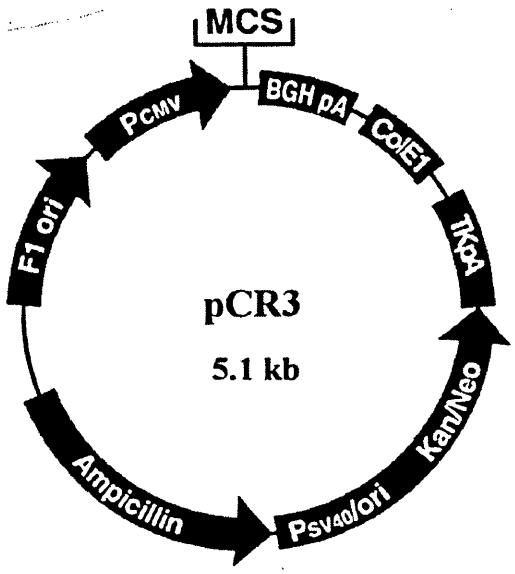



**Figure 2.1** Map of pCR3. Reproduced from the Eukaryotic TA cloning kit instruction manual in which the complete nucleotide sequence can be found.

```
        pCR3seq primer →
5′      AGGTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCG
3′      TCCAGATATATTCGTCTCGAGAGACCGATTGATCTCTTGGGTGACGAATGACCGAATAGC

            T7 primer →              Hind III
5′      AAATTAATACGACTCACTATAGGGAGACCCAAGCTTGGTACCGAGCTCGGATCCACTAGT
3′      TTTAATTATGCTGAGTGATATCCCTCTGGGTTCGAACCATGGCTCGAGCCTAGGTGATCA

                                                        EcoR V
5′      AACGGCCGCCAGTGTGCTGGAATTCGGCTT      AGCCGAATTCTGCAGATATCCATC
3′      TTGCCGGCGGTCACACGACCTTAAGCCGA       TTCGGCTTAAGACGTCTATAGGTAG

                                  Xba I                 ← Sp6 primer
5′      ACACTGGCGGCCGCTCGAGCATGCATCTAGAGGGCCCTATTCTATAGTGTCACCTAAATC
3′      TGTGACCGCCGGCGAGCTCGTACGTAGATCTCCCGGGATAAGATATCACAGTGGATTTAG
```

**Figure 2.2** The map of the multiple cloning site (MCS) of the supplied 5058 bp linear pCR3 vector. Other restriction sites unique to the multiple cloning site include *Kpn* I, *Bam*H I, *Bst*X I, *Eco*R I, *Pst* I, *Not* I, *Xho* I and *Apa* I.

The T7 and Sp6 priming sites encoded by the pCR3 multiple cloning site allow sequencing and PCR amplification of the cloned insert. Including the two 1 bp joints, PCR amplification with the Sp6 and T7 primers appends 171 bp to any insert cloned into the supplied 5058 bp linear pCR3 vector.

Circularised pCR3 was obtained after the loss of both 3′ deoxythymidine overhangs from the linearised vector enabled a blunt-ended self-ligation, which was verified by DNA sequencing. The 5058 bp circularised pCR3 plasmid (in *E. coli* TOP 10F′) was cultured overnight at 37°C in a 200ml volume of LB medium, supplemented with kanamycin to a concentration of 25μg/ml. A 200ml volume of this overnight culture was then used in a large-scale plasmid preparation (Section 2.1.15) yielding a stock solution of 0.2μg/μl. A 2μg quantity of circularised pCR3 vector was digested with *Hin*d III / *Xba* I (Section 2.1.11). The 4946 bp linearised pCR3 vector was separated from the released 104 bp insert (Section 2.1.7) and resuspended in 25μl water, yielding a stock solution of approximately 80ng/μl (25 fmol/μl) *Hin*d III / *Xba* I digested pCR3 (Figure 2.3). Including the two 4 bp restriction site joints, PCR amplification with the Sp6 and T7 primers appends 65 bp to any insert unidirectionally cloned into the *Hin*d III / *Xba* I restriction sites of pCR3. A third primer, denoted pCR3seq, was designed and synthesized to sequence such inserts.

```
        pCR3seq primer →
5′      AGGTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCG
3′      TCCAGATATATTCGTCTCGAGAGACCGATTGATCTCTTGGGTGACGAATGACCGAATAGC

           T7 primer →              Hind III            Xba I
5′      AAATTAATACGACTCACTATAGGGAGACCCA                 CTAGAGGGCCC
3′      TTTAATTATGCTGAGTGATATCCCTCTGGGTTCGA                  TCCCGGG

                  ← Sp6 primer
5′      TATTCTATAGTGTCACCTAAATC
3′      ATAAGATATCACAGTGGATTTAG
```

**Figure 2.3** The map of the multiple cloning site of the linear 4946 bp *Hin*d III / *Xba* I digested pCR3 vector.

### 2.1.4 The pUC119- *Sfi / Not* -$HIS_6$ plasmid

The pUC119- *Sfi / Not* -$HIS_6$ plasmid (Figure 2.4) was obtained from M. W. Robertson, Laboratory of Molecular Biology, Cambridge, UK. The pUC119- *Sfi / Not* -$HIS_6$ plasmid is a modified version of pUC119 (3162bp) (Sambrook *et al.*, 1989)(Vieira and Messing, 1987) that contains a different nucleotide sequence between the *Sph* I and *Eco*R I restriction sites in the multiple cloning site. The pUC119- *Sfi / Not* -$HIS_6$ vector had been designed to unidirectionally clone inserts into the unique *Sfi* I and *Not* I restriction sites, and to express them in the periplasm of *E. coli* under the control of the *lacZ* promoter from *E. coli* (Yanisch-Perron *et al.*, 1985)(Makrides, 1996).

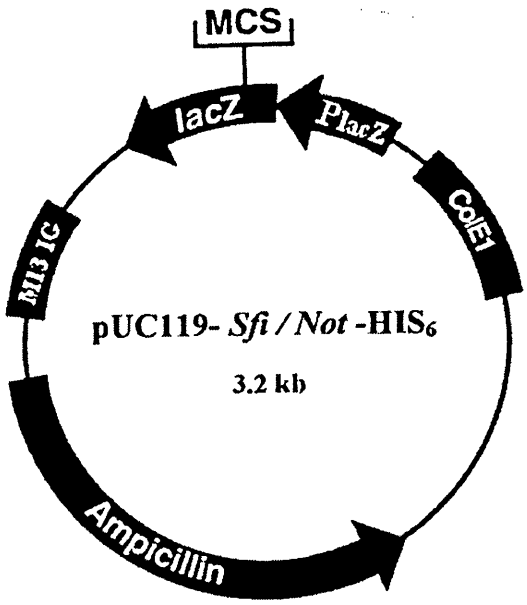



**Figure 2.4** Map of the pUC119- *Sfi / Not* -$HIS_6$ plasmid. Reproduced with alterations from the Eukaryotic TA cloning kit instruction manual.

The pUC119- *Sfi / Not* -$HIS_6$ plasmid was supplied as a circularised vector in *E. coli* TG1 cells (Gibson, 1984)(Sambrook *et al.*, 1989) containing an undefined insert of approximately 500 bp between the *Sfi* I and *Not* I restriction sites. This recombinant was cultured overnight at 37°C in a 300ml volume of 2xYT medium, supplemented with ampicillin to a concentration of 100µg/ml. A 300ml volume of this overnight culture was then used in a large-scale plasmid preparation (Section 2.1.15) yielding a stock solution of 0.62µg/µl. A 3.1µg quantity of circularised pUC119- *Sfi / Not* -$HIS_6$ vector was digested with *Not* I / *Sfi* I (Section 2.1.11) and the 3229 bp linearised vector was separated from the released insert (Section 2.1.7) yielding a stock solution of approximately 135ng/µl (65 fmol/µl) *Not* I / *Sfi* I digested pUC119- *Sfi/Not* -$HIS_6$ (Figure 2.5).

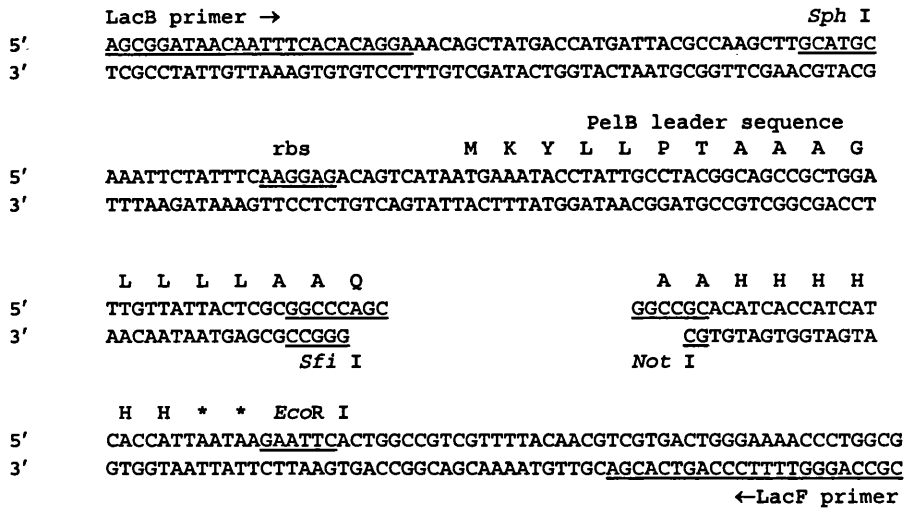



**Figure 2.5** The map of the multiple cloning site of the 3229 bp linear *Sfi* I / *Not* I digested pUC119- *Sfi / Not* -$HIS_6$ vector. Note that the *lacZ* promoter lies upstream from this sequence and that * represents a termination codon.

The LacB and LacF priming sites encoded by the pUC119- *Sfi / Not* -$HIS_6$ multiple cloning site allow sequencing and PCR amplification of the cloned insert. Including the 3 bp and 4 bp restriction site joints, PCR amplification with the LacB and LacF primers appends 223 bp to any insert unidirectionally cloned into the *Sfi* I / *Not* I restriction sites of pUC119- *Sfi / Not* -$HIS_6$.

The multiple cloning site of the *Not* I / *Sfi* I digested pUC119- *Sfi* / *Not* -$HIS_6$ vector encodes a ribosome binding sequence (rbs), the first 18 amino acids of the 22 amino acid (MKYLLPTAAAGLLLLAAQPAMA) PelB leader sequence (Lei *et al.*, 1987)(Better *et al.*, 1988), a six residue histidine tail and a termination codon. For any insert to be successfully expressed in the *E. coli* periplasm it must be transcribed and translated in conjunction with the PelB leader sequence (which encodes the ATG codon for translation initiation). Therefore, the portion of the PelB leader sequence encoded by the plasmid must be completed by the insert to ensure that the PelB leader sequence is functional and that the inserted gene is in frame. Thus, for successful periplasmic protein expression in *E. coli*, the insert must encode a *Sfi* I restriction site, two amino acids (methionine and alanine) to complete the PelB sequence, the gene to be expressed and a *Not* I restriction site (Figure 2.6). Any inserted gene not encoding or encoding a termination codon will be expressed with or without a histidine tail, respectively.

```
        LacB primer→                                              Sph I
5'      AGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTTGCATGC
3'      TCGCCTATTGTTAAAGTGTGTCCTTTGTCGATACTGGTACTAATGCGGTTCGAACGTACG

                                          PelB leader sequence
                rbs              M  K  Y  L  L  P  T  A  A  A  G
5'      AAATTCTATTTCAAGGAGACAGTCATAATGAAATACCTATTGCCTACGGCAGCCGCTGGA
3'      TTTAAGATAAAGTTCCTCTGTCAGTATTACTTTATGGATAACGGATGCCGTCGGCGACCT


         L  L  L  L  A  A  Q  P  A  M  A        A  A  A  H  H  H  H
5'      TTGTTATTACTCGCGGCCCAGCCGGCCATGGCCINSERTGCGGCCGCACATCACCATCAT
3'      AACAATAATGAGCGCCGGGTCGGCCGGTACCGG       CGCCGGCGTGTAGTGGTAGTA
                      Sfi I                      Not I

         H  H  *  *  EcoR I
5'      CACCATTAATAAGAATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCG
3'      GTGGTAATTATTCTTAAGTGACCGGCAGCAAAATGTTGCAGCACTGACCCTTTTGGGACCGC
                                                          ←LacF primer
```

**Figure 2.6** The map of the multiple cloning site of the pUC119- *Sfi* / *Not* -$HIS_6$ vector containing an insert (highlighted) to be expressed in the *E. coli* periplasm. Note that the *lacZ* promoter lies upstream from this sequence.

### 2.1.5 Synthesis of oligonucleotides and quantitation of DNA

All the oligonucleotides used in this work (unless otherwise stated) were synthesized by Perkin-Elmer Ltd, Applied Biosystems Division, Cheshire, UK., and were supplied, unmodified, in 20% (v/v) acetonitrile/water.

The concentrations of the supplied oligonucleotides were verified by measuring their absorbance at 260nm on a Cecil Ce6600 multimode computing UV spectrophotometer (Cambridge, UK), using a matched pair of Hellma quartz cuvettes (path length 10mm), and the following equation provided by Perkin-Elmer (where n represents the number of times the base is represented in the oligonucleotide). Oligonucleotide concentration (pmol/μl) = $[(100x\ A_{260}) \div (1.54nA + 0.75nC + 1.17nG + 0.92nT)]$.

The concentration of a solution of double-stranded DNA was also determined by measuring the absorbance at 260nm. A 50μg/ml dsDNA solution has an absorbance of 1 unit at 260nm (Sambrook *et al*, 1989).

In order to convert μg quantities of oligonucleotides into pmol concentrations, the μg quantity of ssDNA oligonucleotide was multiplied by [(1 000 000)/(length of oligonucleotide x 325)], where 325 represents the average molecular weight of a base in daltons (Perkin-Elmer). For dsDNA molecules the 325 parameter was replaced by 650, which represents the average molecular weight of a base pair in daltons (Perkin-Elmer).

### 2.1.6 Agarose gel electrophoresis

DNA fragments were visualised by horizontal agarose gel electrophoresis using either 1%, 1.5%, 2% or 3% (w/v) agarose gels. Note that all the figures in this work, unless otherwise stated, are of 3% (w/v) agarose gels. UltraPure™ agarose (Life Technologies, Paisley, UK) was dissolved in 50ml 1x TBE buffer (90 mM Tris-borate, 2mM EDTA) by heating. Ethidium bromide (Sigma) was then added to a final

concentration of 0.4 μg/ml. The warm agarose was then poured into a Flowgen (Staffordshire, UK) 10 x 7 cm mini gel unit which had a well-forming comb already clamped into place. The gel was left at room temperature for 30 minutes to set. Once set, the comb and gel spacers were removed and the gel tank was filled with 1x TBE buffer supplemented with 0.4 μg/ml ethidium bromide. Samples, in 1x loading buffer (0.002% (w/v) xylene cyanol FF (Sigma), 6.66% (w/v) sucrose in water), were loaded into the sample wells. This loading buffer contained enough loading dye to only allow visualisation of the actual loading process; not enough loading dye was present to actually see dye migration through the agarose gel. In this way no faint DNA bands could be obscured by the dye. If visualisation of dye migration was desired, a second 1x loading buffer (0.04% (w/v) xylene cyanol FF, 6.66% (w/v) sucrose in water) was used. A constant voltage of 85 V (8.5 V/cm) was applied for between 20 and 45 minutes (depending on the agarose concentration) after which time the size-separated DNA was visualised with UV light.

Alternatively, for greater and more accurate resolution of DNA fragments between 100-250 bp in size, a 3% (w/v) agarose gel in 1x TBE buffer was run on a 20 x 20cm GNA-200 submarine electrophoresis system (Pharmacia) at a constant voltage of 150 V (5 V/cm) for 165 minutes.

### 2.1.7 Purification of DNA from low melting point agarose gels using β-agarase I

Specifically sized DNA fragments were selectively isolated from mixtures of differently sized DNA fragments using either 1% or 1.5% (w/v) low melting point agarose gels. UltraPure™ low melting point agarose (Life Technologies) was dissolved in 1x TAE buffer (40 mM Tris-acetate, 1mM EDTA) and dissolved by heating. Ethidium bromide was added to a final concentrarion of 0.4μg/ml and after the gel was poured it was left in a 4°C cold room for 30 minutes to set. The gel was then run at 40 V (4 V/cm) for 60 minutes at 4°C. The desired DNA band was then excised with a clean razor blade under a long-wave hand-held UV illuminator and

placed in a 1.5ml microtubule (hereafter referred to as an eppendorf). The isolated DNA was then extracted from the excised gel slice using β-agarase I.

The mass of the excised gel slice was determined and it was then heated at 65°C for 10 minutes. When molten, 10% (v/v) 10x β-agarase I buffer (100mM Bis-Tris-HCl [pH 6.5], 10mM EDTA) was added and the reaction was cooled to 40°C. The β-agarase I (1U/μl) was added to a concentration of 1 unit β-agarase per 100μl 1% agarose or 67μl 1.5% agarose. The reaction was incubated at 40°C for 2 hours, after which time sodium acetate (pH 5.5) was added to a final concentration of 0.3M, and the reaction was chilled on ice for 15 minutes. Undigested agarose was pelleted by centrifugation at 13 000 rpm (9440 x g) for 5 minutes on a MSE micro-centaur microfuge. The supernatant was transferred to a fresh eppendorf, two volumes of pre-chilled isopropanol (-20°C) were added and after mixing by inversion, the reaction was then incubated at -20°C for 60 minutes. The precipitated DNA was pelleted by centrifugation at 13000 rpm (9440 x g) for 15 minutes at 4°C. The supernatant was carefully removed and discarded using a Gilson pipetteman, leaving approximately 20μl behind so as not to disturb the DNA pellet. The DNA pellet was then washed by inversion with 500μl 70% (v/v) ethanol, pelleted at 13000 rpm (9440 x g) for 5 minutes (the supernatant being carefully removed as before), dried under vacuum and re-suspended in 20μl of water.

### 2.1.8 Precipitation of DNA with isopropanol

DNA precipitation with isopropanol was used to either concentrate DNA samples or to effect a buffer exchange. Samples to be precipitated, which were less than 100μl in volume, were made up to 100μl with water. DNA was precipitated with the addition of 10% (v/v) 3M sodium acetate (pH 5.5) and two volumes of pre-chilled isopropanol (-20°C), the reaction being mixed by inversion and incubated at -20°C for 60 minutes. The protocol then follows from the same point as described in Section 2.1.7.

### 2.1.9 The polymerase chain reaction (PCR)

The polymerase chain reaction enables the specific amplification of a desired nucleotide sequence, the specificity of the process deriving from the design of the two oligonucleotide primers. Repeated cycles of DNA denaturation, primer annealing, and enzyme-catalysed primer extension, effect the amplification process (McPherson *et al.*, 1992).

PCR reactions were performed on a M. J. Research Inc. PTC-100 programmable thermal controller using the thermostable DNA polymerases *Taq* or $Vent_R$ (obtained from the thermophilic organisms *Thermus aquaticus* and *Thermococcus litoralis*, respectively). $Vent_R$ DNA polymerase exhibits 3′→5′ proofreading exonuclease activity which in part explains why the copying fidelity of $Vent_R$ DNA polymerase ($57 \times 10^{-6}$ errors/bp/duplication [Mattila *et al.*, 1991]) is approximately 5 fold higher than that of *Taq* DNA polymerase ($285 \times 10^{-6}$ errors/bp/duplication [Tindall et al., 1988]).

For a *Taq* PCR reaction, the target DNA template was incubated with the forward and reverse oligonucleotide primers (each at a final concentration 0.5 pmol/µl reaction volume) in 1x *Taq* buffer (11.2mM $(NH_4)_2SO_4$, 46.9mM Tris-HCl (pH 8.8), 0.007% Tween-20, 1.5mM $MgCl_2$), supplemented with each deoxynucleoside 5′ triphosphate (dATP, dCTP, dGTP, dTTP) to a final concentration of 200µM, and BIOTAQ™ *Taq* DNA polymerase to a final concentration of 0.05 units/µl reaction volume, in a final volume of 50µl.

For a $Vent_R$ PCR reaction, the target DNA template was incubated with the forward and reverse oligonucleotide primers (each at a final concentration 2.0 pmol/µl reaction volume) in 1x $Vent_R$ buffer (10mM KCl, 10mM $(NH_4)_2SO_4$, 20mM Tris-HCl (pH 8.8), 2mM $MgSO_4$, 0.1% Triton X-100), supplemented with each deoxynucleoside 5′ triphosphate (dATP, dCTP, dGTP, dTTP) to a final concentration of 300µM, BSA to a final concentration of 0.1µg/µl reaction volume, and $Vent_R$ DNA polymerase to a final concentration of 0.02 units/µl reaction volume, in a final volume of 50µl.

All PCR reactions were overlaid with molecular biology mineral oil (Sigma) to prevent evaporation during the temperature cycling. The precise cycling parameters used for each PCR reaction are listed in the relevant sections, alongside more detailed information on the nature and concentration of the target DNA template. Every PCR reaction was run in conjunction with a negative control, in which the target DNA template was replaced with water. In order to PCR screen recombinant DNA constructs within *E. coli* cells, a sterile toothpick was dipped once into the relevant bacterial colony and subsequently dipped once into a PCR reaction. All such PCR screening reactions used a final reaction volume of 20µl instead of the normal 50µl volume.

### 2.1.11 Restriction endonuclease digestion of DNA

All restriction endonuclease digestion reactions were incubated overnight in a PCR tube, overlaid with mineral oil, in a total volume of 40µl, at the recommended temperature and in the recommended 1x New England Biolabs buffer (NEBuffer), supplied by New England Biolabs. The 1x buffer compositions are as follows; NEBuffer 1: 10mM Bis Tris Propane-HCl, 10mM $MgCl_2$, 1mM DTT (pH 7.0); NEBuffer 2: 10mM Tris-HCl, 10mM $MgCl_2$, 50mM NaCl, 1mM DTT (pH 7.9); NEBuffer 3: 50mM Tris-HCl, 10mM $MgCl_2$, 100mM NaCl, 1mM DTT (pH 7.9); NEBuffer 4: 20mM Tris-acetate, 10mM magnesium acetate, 50mM potassium acetate, 1mM DTT (pH 7.9). Bovine serum albumin was also added to a final concentration of 100µg/ml when recommended.

DNA was digested simultaneously with *Hin*d III and *Xba* I as the reaction requirements for the two enzymes were compatible. DNA was digested with 20 units *Hin*d III and 20 units *Xba* I at 37°C overnight in 1x NEBuffer 2 supplemented with 100µg/ml BSA, in a final volume of 40µl. Both enzymes were heat inactivated at 65°C for 20 minutes.

DNA was digested with 10 units *Mnl* I at 37°C overnight in 1x NEBuffer 2 supplemented with 100µg/ml BSA, in a final volume of 40µl.

DNA was digested sequentially with *Not* I and *Sfi* I as the reaction requirements for the two enzymes were incompatible. DNA was digested with 20 units *Not* I at 37°C overnight in 1x NEBuffer 3 supplemented with 100µg/ml BSA, in a final volume of 40µl. The *Not* I was then heat inactivated at 65°C for 20 minutes. The DNA was then digested with 10 units *Sfi* I at 50°C overnight. The restriction enzyme *Sfi* I cannot be heat inactivated (as determined by the manufacturer). The DNA was separated from the active *Sfi* I, as detailed in Section 2.1.7. Although 1x NEBuffer 2 is the recommended buffer for *Sfi* I, the enzyme retains 10% activity in NEBuffer 3 (as determined by the manufacturer).

### 2.1.12 Ligation of inserts into the pCR3 and pUC119- *Sfi* / *Not* -$HIS_6$ plasmids

Insert and plasmid DNA was ligated with 200 units T4 DNA ligase at 16°C overnight in 1x T4 DNA ligase buffer (50mM Tris-HCl, 10mM $MgCl_2$, 10mM DTT, 1mM ATP, 50µg/ml BSA (pH7.8)), in a final volume of 10µl. The reaction took place in a PCR tube at the recommended temperature (as determined by the manufacturer). Afterwards the T4 DNA ligase was heat inactivated at 65°C for 10 minutes. The precise amount of plasmid used in each ligation reaction, alongside detailed information on the actual nature and amount of the insert being cloned, can be found in the relevant section.

The strategy of choice when cloning inserts into plasmid vectors, which was used throughout this work, was to digest the vector and insert with two different restriction enzymes, producing DNA fragments with noncomplementary overhangs, so that on ligation the insert is unidirectionally cloned into the vector, the restriction sites are preserved, and a low background of non-recombinant vector is obtained. Although it is recommended to use between a 1:1 or 1:2 vector to insert molar ratio in a ligation reaction involving cohesive termini (Sambrook *et al.*, 1989), in practice it is often not

possible to know the exact concentration of the insert DNA, which in this work was usually estimated by visualisation after agarose gel electrophoresis (Section 2.1.6).

### 2.1.13 Preparation and transformation of competent *E. coli* (TOP 10F')

Competent cells were prepared and transformed using a variation of the procedures detailed in Sambrook *et al.* (1989) and Cohen *et al.* (1972). A 5µl volume of *E. coli* TOP 10F' glycerol stock (Section 2.1.2) was inoculated into 2xYT (5ml) with no antibiotic supplements and incubated overnight in a rotary shaker (200 cycles/minute) at 37°C. A 50ml volume of 2xYT media, in a 500ml flask, was then inoculated with 5µl of this overnight culture and incubated, with shaking, at 37°C until the absorbance of the culture at 550nm was 0.4 units. The cells were then centrifuged at 1000 x g for 5 minutes and the cell pellet resuspended in a 25ml volume of ice-cold 50mM $CaCl_2$. After 20 minutes incubation on ice, the cells were pelleted (as before) and resuspended in a 5ml volume of ice-cold 50mM $CaCl_2$/15% (v/v) glycerol solution. Cell aliquots of 200µl were dispensed into chilled 0.5ml eppendorfs. The competent cells were then snap-frozen in an acetone/dry-ice bath and stored at -70°C.

To transform *E. coli* TOP 10F' a frozen aliquot of competent cells was thawed on ice and a 100µl volume of competent cells was incubated with a 3µl volume of ligation reaction (Section 2.1.12) on ice for 2 hours. The cells were then heat-shocked at 42°C for 2 minutes and returned to ice for a further 2 minutes. A 400µl volume of 2xYT media was added and the cells were incubated, with shaking, at 37°C for 60 minutes to allow cell recovery. Aliquots of 200µl were then spread onto 2xYT agar plates, supplemented with ampicillin to a concentration of 50 µg/ml (for pCR3 plasmids) or 100 µg/ml (for pUC119- *Sfi* / *Not* -$HIS_6$ plasmids). The plates were then inverted and incubated at 37°C overnight.

### 2.1.14 Small-scale preparation of plasmid DNA

Small-scale purifications of plasmid DNA, commonly known as minipreps, were achieved using the Wizard™ minipreps DNA purification system (Promega). *E. coli* cells, carrying the plasmid of interest, were incubated in a 5ml volume of 2xYT media, supplemented with ampicillin to a concentration of 50 μg/ml (for pCR3) or 100 μg/ml (for pUC119- *Sfi / Not* -$HIS_6$), at 37°C overnight. A 3ml volume of this overnight culture was treated with the Wizard minipreps kit as described in the manufacturer's instructions.

### 2.1.15 Large-scale preparation of plasmid DNA

Large-scale purifications of plasmid DNA, commonly known as maxipreps, were achieved using the Wizard™ maxipreps DNA purification system (Promega). *E. coli* cells, carrying the plasmid of interest, were incubated in a 200-300ml volume of 2xYT or LB media (as specified), supplemented with ampicillin to a concentration of 50 μg/ml (for pCR3) or 100 μg/ml (for pUC119- *Sfi / Not* -$HIS_6$), at 37°C overnight. A 200-300ml (as specified) volume of this overnight culture was treated with the Wizard maxipreps kit as described in the manufacturer's instructions. The concentration of the eluted plasmid DNA was determined (Section 2.1.5) and the plasmid DNA was stored at -20°C.

### 2.1.16 Automated DNA sequencing and sequence analysis

Ultra-pure plasmid DNA, for use in automated DNA sequencing reactions, was obtained using the S.N.A.P. miniprep kit (Invitrogen). *E. coli* cells carrying the plasmid DNA to be sequenced were grown at 37°C overnight in a 4ml volume of 2xYT media, supplemented with ampicillin to a concentration of 50 μg/ml (for pCR3) or 100 μg/ml (for pUC119- *Sfi / Not* -$HIS_6$). A 3ml volume of this overnight culture was treated with the S.N.A.P. kit as described in the manufacturer's instructions. A

6µl volume of the S.N.A.P. miniprep (approximately 1500ng plasmid DNA) was made up to a final volume of 12µl with water and 3.2 pmol of sequencing primer. This reaction mix was then submitted for automated DNA sequencing, based upon the dideoxy-sequencing method (Sanger and Coulson, 1977), in an ABI PRISM 377 DNA sequencer (Perkin Elmer, Foster City, USA) at the University of Bath. The resulting DNA sequences were analysed using the SeqEd™ (sequence editor) program, version 1.0.3. (Applied Biosystems, a division of Perkin Elmer, Foster City, USA) on an Apple Macintosh microcomputer.

### 2.1.17 Non-denaturing polyacrylamide gel electrophoresis

Non-denaturing polyacrylamide gels were used for the visualisation of ssDNA oligonucleotides and for the separation and purification of small dsDNA fragments. Polyacrylamide gels of 12% and 20% (w/v) polyacrylamide effectively separate dsDNA fragments between 40-200 and 6-100 bp, respectively (Sambrook *et al.*, 1989).

A 20ml volume of non-denaturing acrylamide gel solution, 12% or 20% (w/v) acrylamide (Protogel) in 1x TBE buffer (90mM Tris-Borate, 2mM EDTA), was polymerised by the addition of 14mg ammonium persulphate (from a 10% (w/v) solution) and a 7µl volume of TEMED. A 4ml volume of this gel solution was pipetted into a BIO-RAD Mini Protean II gel apparatus. A well-forming comb was positioned in the gel, which was then left to set at room temperature for 60 minutes. The gel was then placed in an electrophoresis tank and the upper and lower tanks were filled with 1x TBE. The well-forming combs were carefully removed and any unpolymerised acrylamide was washed from the sample wells with 1x TBE. The DNA samples were then loaded in 1x loading buffer (6.66% (w/v) sucrose in water) using a Hamilton syringe (Hamilton, Reno, USA). Samples were electrophoresed at 80 V (8 V/cm) for 120 minutes. No xylene cyanol FF dye was in the loading buffer so as not to obscure any faint DNA bands on the polyacrylamide gel, although loading buffer containing xylene cyanol FF (0.04% (w/v) xylene cyanol FF, 6.66% (w/v)

sucrose in water) was run alongside the DNA samples in order to visually follow the course of the electrophoresis. Note that the xylene cyanol FF dye comigrates with dsDNA fragments of 70 bp and 45 bp on 12% and 20% acrylamide gels, respectively (Sambrook *et al.*, 1989).

After electrophoresis the DNA was visualised by UV "shadowing". The acrylamide gel was carefully transferred onto a piece of clingfilm covering a fluorescent TLC silica plate. A hand-held 254/366nm UV illuminator was shone onto the gel. The DNA present in the acrylamide gel was visualised as a black band (or "shadow") on the fluorescent TLC plate.

DNA was recovered from non-denaturing polyacrylamide gels using Micropure-0.22 inserts and Microcon-10 microconcentrators (Amicon). The DNA band of interest was excised from the polyacrylamide gel with a clean razor blade, placed into a 1.5ml eppendorf and ground into small pieces with a 1000µl disposable pipette tip. A 200µl volume of water was added and the gel slurry was incubated at 37°C for 2 hours. The gel slurry was transferred to a Micropure-0.22 insert which was inside a Microcon-10 microconcentrator that had been inserted into a 1.5ml eppendorf. A 100µl volume of water was used to rinse the eppendorf that had contained the gel slurry; this was then added to the gel slurry already in the Micropure-0.22 insert. Centrifugation at 13000 rpm (9440 x g) for 20 minutes left the acrylamide pieces in the Micropure-0.22 insert, the DNA above the membrane in the Microcon-10 microconcentrator (which retains ssDNA above 30 nt and dsDNA above 20 bp) and the aqueous solution in the eppendorf. The Micropure-0.22 insert was discarded and in order to remove any salts from the DNA, a 400µl volume of water was added to the Microcon-10 which was then centrifuged at 13000 rpm (9440 x g) for 25 minutes. To recover the DNA, the Microcon-10 was inverted and placed into a new eppendorf. Centrifugation at 4000 rpm (1000 x g) for 3 minutes eluted the purified DNA in approximately 10µl water.

### 2.1.18 Phosphorylation of oligonucleotides

Single-stranded DNA oligonucleotides were phosphorylated at the 5′ hydroxyl terminus using T4 polynucleotide kinase. A 300pmol quantity of ssDNA oligonucleotide was phosphorylated with 10 units T4 polynucleotide kinase at 37°C for 60 minutes in 1x T4 DNA ligase buffer (Section 2.1.12), in a final volume of 50µl. The T4 polynucleotide kinase was then heat inactivated at 65°C for 20 minutes. Stocks of phosphorylated oligonucleotides, at a concentration of 6pmol/µl, were stored at -20°C. T4 polynucleotide kinase is supplied with its own buffer, 1x T4 polynucleotide kinase buffer (70mM Tris-HCl [pH 7.6], 10mM $MgCl_2$, 5mM DTT) which must be supplemented with 1mM ATP, although it exhibits 100% activity in 1x T4 DNA ligase buffer (as determined by the manufacturer). Phosphorylating the oligonucleotides in 1x T4 DNA ligase buffer allowed the stocks to be used directly in dsDNA block-forming reactions and subsequent gene-assembling ligation reactions (Section 2.1.19).

### 2.1.19 Gene-assembling ligation reactions

All the dsDNA blocks were prepared by annealing pairs of complementary oligonucleotides by lyophilisation. An equimolar amount (from 1-60pmol, as specified) of the two oligonucleotides constituting a particular block was incubated in a 0.5ml eppendorf in 1x T4 DNA ligase buffer (Section 2.1.12), in a final volume of 25µl. The eppendorf was sealed and the lid was pierced with a needle. The reaction was frozen in an acetone/dry-ice bath for 60 seconds and quickly transferred to a pre-chilled (-20°C) 500ml round-bottomed flask. The flask was then attached to an I.E.C. Lyoprep-3000 freeze-drier (International Equipment Company, Dunstable, U.K.), using high pressure vacuum grease to ensure an air-tight fit. The samples were lyophilised for 90 minutes under vacuum (0.05-0.15mbar). The flask was then quickly transferred to a 4°C cold room and the lyophilised blocks were re-constituted in 1x T4 DNA ligase buffer with a 25µl volume of water (pre-chilled to 4°C).

The dsDNA blocks were subsequently mixed together (as specified), in a two stage ligation reaction, in order to assemble the genes encoding the peptide library. In the primary ligation reaction all the resuspended TGT, TGC, TGT[ ] and TGC[ ] propagating blocks were quickly transferred to a 1.5ml eppendorf, and thoroughly mixed together. T4 DNA ligase was then added to a final concentration of 4 units/μl reaction volume and the reaction was incubated at 4°C for 1 hour. The secondary ligation reaction involved the addition of the N and C terminating blocks to the primary ligation reaction with a further 200 units T4 DNA ligase (preserving the T4 DNA ligase concentration of 4 units/μl reaction volume). Before addition, the two resuspended terminating blocks had been kept separate at 4°C. The secondary ligation reaction was then incubated at 4°C for 2 hours, after which time the T4 DNA ligase was heat inactivated at 65°C for 10 minutes.

**N.B.** The 5′-phosphorylated oligonucleotide stock solutions (Section 2.1.18), at a concentration of 6pmol/μl, contain 0.01μl (v/v) glycerol/μl reaction volume, as the T4 polynucleotide kinase was stored in 50% (v/v) glycerol. A residual volume of glycerol greater than 0.05μl prevents complete lyophilisation of the blocks. Consequently, the maximum amount of 5′-phosphorylated oligonucleotide stock solution that can be fully lyophilised is 5μl; enough to form 15pmol of dsDNA block. If higher amounts of block are required, oligonucleotides that have been supplied phosphorylated must be used or several lyophilisations of the same block must be added to the ligation reaction.

# CHAPTER 3

# DETERMINATION OF THE STRATEGY FOR CONSTRUCTING THE DNA LIBRARY

## 3.0 INTRODUCTION

The aim of this project is to determine a means of constructing a DNA library encoding a population of variable-length peptides ( $\leq 30$ amino acids in length) which contain an even number of randomly distributed cysteine residues (2, 4, 6 or 8).

In a modification of an idea originally proposed by S. A. Kauffman and M. Ballivet (Kauffman, 1993), the DNA fragments (referred to as genes) encoding the peptide repertoire are to be assembled by the T4 DNA ligase-catalysed polymerisation of small, sticky-ended dsDNA building blocks.

This chapter describes three strategies for constructing the DNA library. In each one there are four classes of dsDNA building block, two which propagate polymerisation by concatenating alternately (the propagating blocks) and two which terminate polymerisation (the terminating blocks). The central point of each strategy is the overhang scheme, which in each case favours the assembly of genes encoding an even number of cysteines, thus the corresponding peptides have the ability to form an integral number of disulphide bonds. Note that assembled genes encoding 2, 4, 6, 8 ... cysteines are denoted constructs I, II, III, IV ..., respectively, and that the length of each encoded peptide is defined as the number of amino acids between its outermost cysteines inclusively.

In the first strategy (Figure 3.1) the N, NNT, NNG and C block classes bear one nucleotide 3′ overhangs. The NNT and NNG propagating blocks are named according to the random amino acid codon represented at their 3′ overhang, whilst the

N and C terminating blocks are named according to the peptide terminus they encode. The overhang scheme of this strategy ensures that the joint between two ligated blocks encodes either 15 (rendered by NNT) or 13 (rendered by NNG) random amino acids. The cysteine residues are encoded within the NNT and NNG blocks, whilst the spacers in these blocks encode a variable number of amino acids, allowing the length of the encoded peptide to vary and the random distribution of the cysteines.

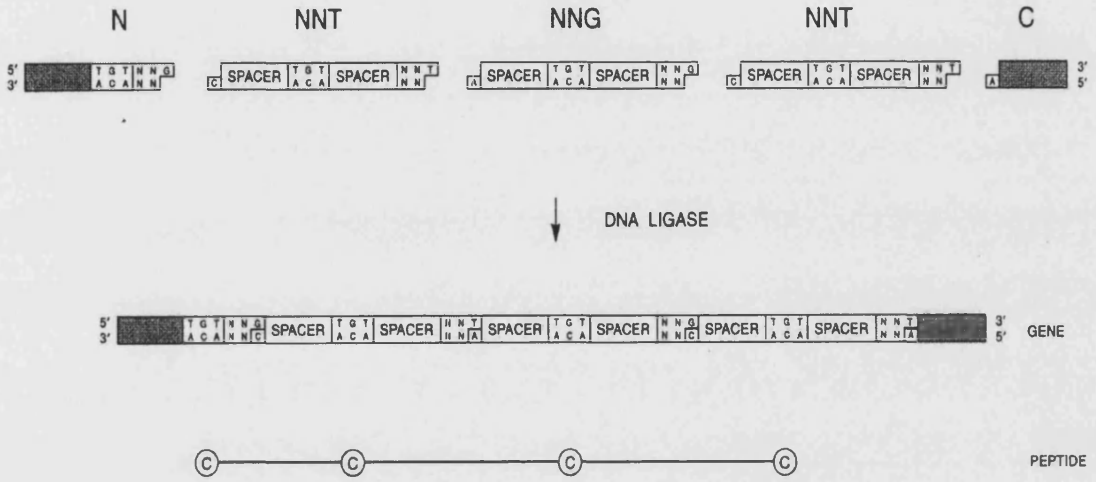



**Figure 3.1** First strategy for constructing a DNA library encoding a population of variable length peptides containing an even number of randomly distributed cysteines. The one nucleotide 3′ overhang scheme utilised by the N, NNT, NNG and C blocks ensures that the joint between two ligated blocks encodes either 15 (rendered by NNT) or 13 (rendered by NNG) random amino acids, and favours the assembly of genes encoding an even number of cysteines. The cysteine residues are encoded within the blocks, and the spacers encode 0, 1, ... random amino acids, allowing the length of the encoded peptide to vary and the random distribution of the cysteines. The shaded regions contain restriction sites for vector incorporation, and an encircled C represents a cysteine residue. The assembled gene shown encodes 4 cysteine residues, but genes encoding 2, 6, 8 ... cysteine residues may also be formed. Note that one of the terminating blocks must encode a cysteine residue for the assembled genes to encode an even number of cysteines, and that the encoded peptides cannot contain two or more contiguous cysteines (unless they have been introduced by the random amino acids).

In the second strategy (Figure 3.2) the N, TGT, TGC and C block classes bear one nucleotide 3′ overhangs. The TGT and TGC propagating blocks are named according to the cysteine codon represented at their overhang, whilst the N and C terminating blocks are named according to the peptide terminus they encode. The overhang scheme of this strategy, which hinges upon the fact that there are two codons for cysteine in the genetic code (5′ UGU 3′ and 5′ UGC 3′), ensures that the joint between two ligated blocks encodes a cysteine residue. The spacers in the blocks encode a variable number of amino acids, allowing the length of the encoded peptide to vary and the random distribution of the cysteines.

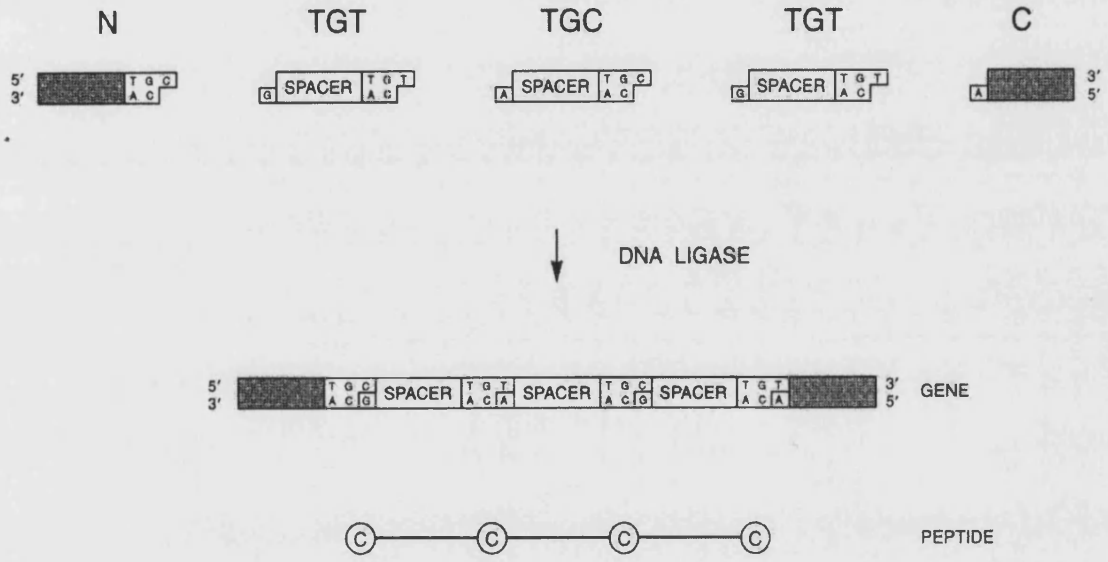



**Figure 3.2** Second strategy for constructing a DNA library encoding a population of variable length peptides containing an even number of randomly distributed cysteines. The one nucleotide 3′ overhang scheme utilised by the N, TGT, TGC and C blocks ensures that the joint between two ligated blocks encodes a cysteine residue, and favours the assembly of genes encoding an even number of cysteines. The spacers encode 0, 1, … amino acids, allowing the length of the encoded peptide to vary and all arrangements of cysteines. The shaded regions contain restriction sites for vector incorporation, and an encircled C represents a cysteine residue. The assembled gene shown encodes 4 cysteine residues, but genes encoding 2, 6, 8 … cysteine residues may also be formed.

In the third strategy (Figure 3.3) the N, TGT, TGC and C blocks bear three nucleotide 5′ overhangs. This strategy is essentially the same as the second strategy (Figure 3.2), the only difference being in the length and orientation of the block overhangs.

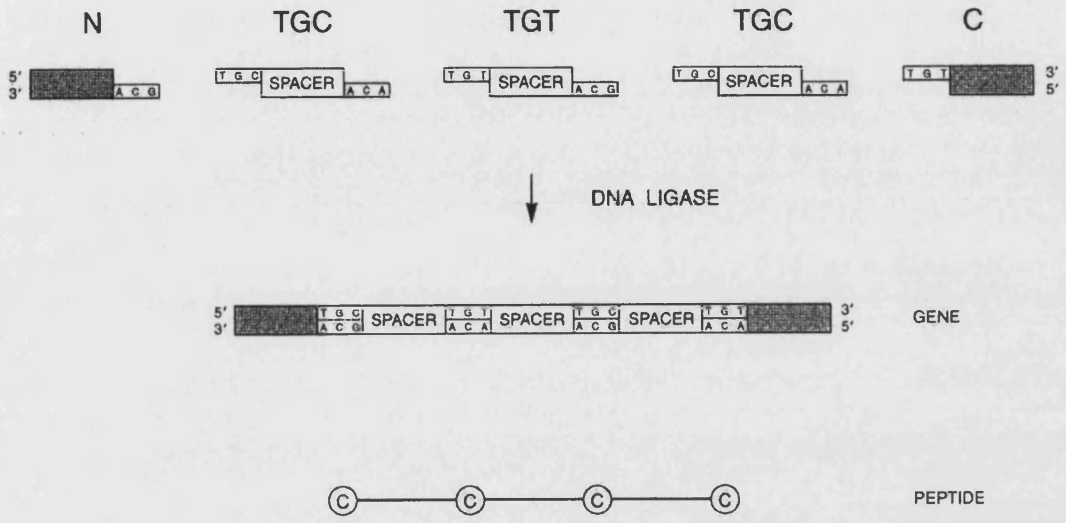



**Figure 3.3** Third strategy for constructing a DNA library encoding a population of variable length peptides containing an even number of randomly distributed cysteines. The three nucleotide 5′ overhang scheme utilised by the N, TGT, TGC and C blocks ensures that the joint between two ligated blocks encodes a cysteine residue, and favours the assembly of genes encoding an even number of cysteines. The spacers encode 1, 2, ... amino acids, allowing the length of the encoded peptide to vary and most arrangements of cysteines. The shaded regions contain restriction sites for vector incorporation, and an encircled C represents a cysteine residue. The assembled gene shown encodes 4 cysteine residues, but genes encoding 2, 6, 8 ... cysteine residues may also be formed. Note that a propagating block with a spacer encoding 0 amino acids is non-existent; a limitation discussed in Chapter 4.

All the dsDNA blocks in this chapter were procured by digesting dsDNA cassettes with restriction endonucleases (Figure 3.4). The first strand of each cassette was a chemically synthesized oligonucleotide, the second strand was formed using a suitable oligonucleotide primer and the Klenow fragment of *E. coli* DNA polymerase I (Oliphant *et al.*, 1986). As it is possible to use either non-degenerate or degenerate oligonucleotides as the first strand of the DNA cassette, this method allows the production of blocks encoding either fixed or random amino acids. Note that random

amino acids can be rendered by the NN(T/G) codon which encompasses all 20 amino acids and an amber stop codon, which can be avoided by using a *supE* suppressor strain of *E. coli* (Sambrook *et al.*, 1989). The cassettes are designed so that the blocks are housed between cut sites for the restriction endonucleases *Mnl* I or *Ear* I, both of which cut at positions removed from their recognition sequence, liberating the blocks with one nucleotide 3′ overhangs and three nucleotide 5′ overhangs, respectively (Figure 3.5).

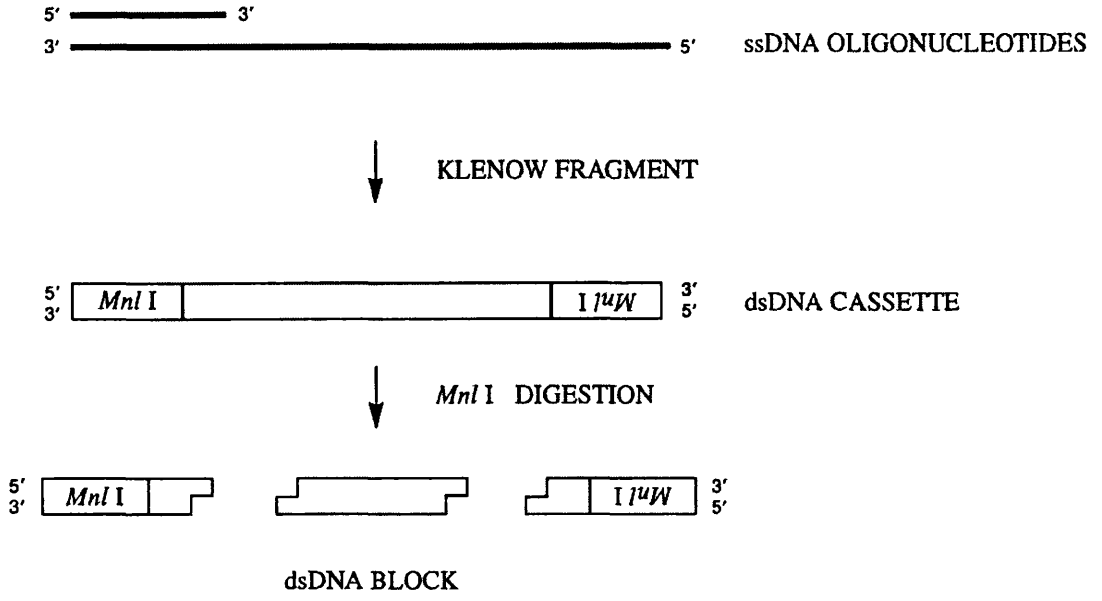



**Figure 3.4** Production of the dsDNA blocks. A chemically synthesized oligonucleotide is made double-stranded to form a DNA cassette using a suitable oligonucleotide primer and the Klenow fragment. Digestion of the dsDNA cassette with a restriction enzyme whose recognition sites are encoded by the dsDNA cassette (in this case *Mnl* I) liberates a dsDNA block (in this case one with one nucleotide 3′ overhangs) along with extraneous digestion products.

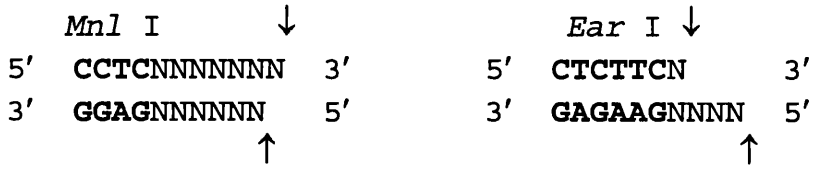



**Figure 3.5** The recognition sequences (in bold text) and cut sites (denoted by the arrows) for the restriction endonucleases *Mnl* I and *Ear* I. Note that N represents any nucleotide.

The N, TGT, TGC and C blocks utilising the one nucleotide overhang scheme (Figure 3.2) are housed in dsDNA cassettes between cut sites for the restriction enzyme *Mnl* I, which releases the blocks with one nucleotide 3′ overhangs. When it was decided to increase the length of the overhangs to three nucleotides, it transpired that no commercially available restriction endonuclease could release the blocks with three nucleotide 3′ overhangs. Consequently, the three nucleotide overhangs were transposed to the 5′ end of the N, TGT, TGC and C blocks (Figure 3.3), which were housed in dsDNA cassettes between cut sites for the restriction enzyme *Ear* I. As a result, the TGT and TGC blocks had effectively swapped places in the strategy (Figures 3.2 and 3.3).

## 3.1 EXPERIMENTAL DESIGN

### 3.1.1 Formation and design of the NNT1 and NNG1 blocks

The 11 base Short A1 and Short A2 oligonucleotides were designed to anneal to the 40 base A1 and 47 base A2 oligonucleotides, respectively (Section 3.2.1), so that treatment with the Klenow fragment would make oligonucleotides A1 and A2 double-stranded. These dsDNA cassettes could then be digested with *Mnl* I to release the 8 bp NNT1 and 11 bp NNG1 blocks (Figure 3.6), respectively, and extraneous DNA fragments of 15 bp and 17 bp, respectively. The NNT1 and NNG1 blocks belonged to the first strategy (Figure 3.1).

```
NNT1        5′   NNKTGTNNT    3′
            3′  CNNMACANN     5′

NNG1        5′   NNKTGTNNKNNG    3′
            3′  ANNMACANNMNN     5′
```

**Figure 3.6** Design of the 8 bp NNT1 and 11 bp NNG1 blocks. The redundant nucleotides are designated N (G/C/A/T), K (G/T), M (A/C).

### 3.1.2 Formation and design of the N1, TGT1, TGC1 and C1 blocks

The 12 base Uniprime B oligonucleotide was designed to anneal to the 57 base B1, 62 base B2, 62 base B3 and 55 base B4 oligonucleotides (Section 3.2.1) so that treatment with the Klenow fragment would make oligonucleotides B1-4 double-stranded. These dsDNA cassettes could then be digested with *Mnl* I to release the 42 bp N1, 32 bp TGT1, 32 bp TGC1 and 40 bp C1 blocks (Figure 3.7), respectively, and extraneous DNA fragments of 14 bp in each case. The N1, TGT1, TGC1 and C1 blocks were intentionally large so that they could be easily isolated by polyacrylamide gel electrophoresis. The N1, TGT1, TGC1 and C1 blocks belonged to the second strategy (Figure 3.2).

```
                    Sfi I
N1          5'  GTGGGCCCAGCCGGCCATGGCAAGGGCCTGCCATGAGCTCTGC   3'
            3'  CACCCGGGTCGGCCGGTACCGTTCCCGGACGGTACTCGAGAC   5'

TGT1        5'   TGGGCATGCAAATATAGGGCGTGCGAATCATGT   3'
            3'  GACCCGTACGTTTATATCCCGCACGCTTAGTAC    5'

TGC1        5'   GCCACCTGTCATACGCACGAACCTATTGGATGC   3'
            3'  ACGGTGGACAGTATGCGTGCTTGGATAACCTAC    5'

                                    Not I
C1          5'   CTAGCAATAAGGGAAGGGGCGGGGGCGGCCGCTAAACTAT   3'
            3'  AGATCGTTATTCCCTTCCCCGCCCCCGCCGGCGATTTGATA   5'
```

**Figure 3.7** Nucleotide sequences of the N1, TGT1, TGC1 and C1 dsDNA blocks.

The N1 and C1 blocks contain *Sfi* I and *Not* I restriction sites, respectively, allowing the assembled genes to be unidirectionally cloned into pUC119- *Sfi* I/*Not* I -$HIS_6$ (Section 2.1.4). This prevents the cysteine codons from flipping to become either threonine or alanine.

### 3.1.3 Formation and design of the N2, TGT2, TGC2 and C2 blocks

The 11 base Uniprime C oligonucleotide was designed to anneal to the 58 base C1, 66 base C2, 66 base C3 and 58 base C4 oligonucleotides (Section 3.2.1) so that treatment with the Klenow fragment would make oligonucleotides C1-4 double-stranded. These dsDNA cassettes could then be digested with *Ear* I to release the 40 bp N2, 32 bp TGT2, 32 bp TGC2 and 40 bp C2 blocks (Figure 3.8), respectively, and extraneous DNA fragments of 15 bp in each case. The N2, TGT2, TGC2 and C2 blocks belonged to the third strategy (Figure 3.3).

```
                    Sfi I
N2        5'  GTGGGCCCAGCCGGCCATGGCAAGGGCCTGCCATGAGCTC      3'
          3'  CACCCGGGTCGGCCGGTACCGTTCCCGGACGGTACTCGAGACG   5'

TGT2      5'  TGTGCCACCTGTCATACGCACGAACCTATTGGA      3'
          3'     CGGTGGACAGTATGCGTGCTTGGATAACCTACG   5'

TGC2      5'  TGCTGGGCATGCAAATATAGGGCGTGCGAATCA      3'
          3'     ACCCGTACGTTTATATCCCGCACGCTTAGTACA   5'

                                        Not I
C2        5'  TGTCTAGCAATAAGGGAAGGGGCGGGGGCGGCCGCTAAACTAT   3'
          3'     GATCGTTATTCCCTTCCCCGCCCCCGCCGGCGATTTGATA   5'
```

**Figure 3.8** Nucleotide sequences of the N2, TGT2, TGC2 and C2 dsDNA blocks.

The N2, TGC2, TGT2 and C2 blocks were redesigned versions of the N1, TGC1, TGT1 and C1 blocks (Section 3.1.2) in which the one nucleotide 3′ overhang scheme (Figure 3.2) had been replaced by the three nucleotide 5′ overhang scheme (Figure 3.3).

## 3.2 MATERIALS AND METHODS

### 3.2.1 Synthesis and design of oligonucleotides

The nucleotide sequences of the oligonucleotides used to form the dsDNA cassettes in this chapter are detailed (Figure 3.9). The A1, B1, Short A2, Short B2, N1, TGT1,

TGC1, C1 and Uniprime 1 oligonucleotides were synthesized by Severn Biotech Ltd., Unit 2, Park Lane, Kidderminster, Worcs., DY11 6TJ. The Severn Biotech oligonucleotides had been HPLC purified and were supplied unmodified and lyophilised, the oligonucleotides were subsequently resuspended in water. The N1, TGT1, TGC1 and C1 oligonucleotides had also been purified by polyacrylamide gel electrophoresis by Severn Biotech.

```
Short A1    (11 nt)      5' GCTGTCCTCGC 3'

Short A2    (11 nt)      5' CTGCTGTCCTC 3'

Uniprime B  (12 nt)      5' CGTCCCTCGAGT 3'

Uniprime C  (11 nt)      5' GCACGTCCCTC 3'

A1  (40 nt)    5' GCGTCCCTCGAGTGCGNNKTGTNNTCCGTGCGAGGACAGC 3'

A2  (47 nt)    5' GGGCGTCCCTCGAGTGCTNNKTGTNNKNNGCCGTGCGAGGACAGCAG 3'

B1  (57 nt)    5' GTGGGCCCAGCCGGCCATGGCAAGGGCCTGCCATGAGCTCTGCGCACTCGAGGGACG 3'

B2  (62 nt)    5' GTGTCCTCGCACGGACATGATTCGCACGCCCTATATTTGCATGCCCAGGCACTCGAGGGACG 3'

B3  (62 nt)    5' GTGTCCTCGCACGGGCATCCAATAGGTTCGTGCGTATGACAGGTGGCAGCACTCGAGGGACG 3'

B4  (55 nt)    5' ATAGTTTAGCGGCCGCCCCCGCCCCTTCCCTTATTGCTAGAGCACTCGAGGGACG 3'

C1  (58 nt)    5' GTGGGCCCAGCCGGCCATGGCAAGGGCCTGCCATGAGCTCTGCAGAAGAGGGACGTGC 3'

C2  (66 nt)    5' ATCGTGTCCTCTTCCGCATCCAATAGGTTCGTGCGTATGACAGGTGGCACAAGAAGAGGGACGTGC 3'

C3  (66 nt)    5' ATCGTGTCCTCTTCCACATGATTCGCACGCCCTATATTTGCATGCCCAGCAAGAAGAGGGACGTGC 3'

C4  (58 nt)    5' ATAGTTTAGCGGCCGCCCCCGCCCCTTCCCTTATTGCTAGACAAGAAGAGGGACGTGC 3'
```

**Figure 3.9** Nucleotide sequences of the oligonucleotides used to form the dsDNA cassettes. The redundant nucleotides are designated N (G/C/A/T) and K (G/T).

### 3.2.2 Formation of dsDNA cassettes

The ssDNA oligonucleotide representing the first strand of the cassette and the short ssDNA oligonucleotide primer, were incubated together with a 6µl volume of 10x *E. coli* DNA polymerase I buffer, in a final volume of 52µl. Detailed information on the nature, amount and concentration of the oligonucleotides can be found in the relevant sections. The oligonucleotides were initially denatured at 94°C for 1 minute, before

being annealed together at 37°C for 60 minutes and cooled to room temperature for 30 minutes. The reaction was supplemented with 10 units Klenow fragment of *E. coli* DNA polymerase I and each deoxynucleoside 5′ triphosphate (dATP, dCTP, dGTP, dTTP) to a final concentration of 250μM, and incubated at 37°C for 2 hours in 1x *E. coli* polymerase I buffer (10mM Tris-HCl, 5mM $MgCl_2$, 7.5mM DTT (pH 7.5)), in a final volume of 60μl. The Klenow fragment was subsequently heat inactivated at 75°C for 20 minutes.

### 3.2.3 Limitations of isolating dsDNA blocks from dsDNA cassettes

The 40 base A1 and 47 base A2 oligonucleotides (Section 3.2.1) were made into dsDNA cassettes (Section 3.2.2) by incubating 11μg of A1 (845pmol) and A2 (720pmol) with 5μg of the Short A1 (1400pmol) and Short A2 (1400pmol) oligonucleotides (Section 3.2.1), respectively. A 30μl volume of each Klenow reaction was electrophoresed on a 20% non-denaturing polyacrylamide gel, and the 40 bp A1 and 47 bp A2 dsDNA cassettes (Section 3.1.1) were isolated and re-suspended in 10μl water (Section 2.1.17). An 8μl volume of each cassette was digested with 10 units *Mnl* I at 4°C overnight in 1xNEBuffer 2 (Section 2.1.11) supplemented with 100μg/ml BSA, in a final volume of 40μl (we found that *Mnl* I digestion does occur at 4°C). The *Mnl* I was not subsequently heat inactivated, so that the small DNA fragments would not be denatured. A 30μl volume of each digestion reaction was electrophoresed on a 20% non-denaturing polyacrylamide gel (Section 2.1.17) in order to purify the released 8 bp NNT1 and 11 bp NNG1 blocks (Section 3.1.1).

### 3.2.4 Assessment of the second strategy for constructing the DNA library

The 57 base B1, 62 base B2, 62 base B3 and 55 base B4 oligonucleotides (Section 3.2.1) were made into dsDNA cassettes (Section 3.2.2) by incubating 12.4μg of B1 (670pmol), B2 (615pmol), B3 (615pmol) and B4 (690pmol) with 20μg of the Uniprime B (5130pmol) oligonucleotide (Section 3.2.1).

A 30µl volume of each Klenow reaction was electrophoresed on a 20% non-denaturing polyacrylamide gel and the 57 bp B1, 62 bp B2, 62 bp B3 and 55 bp B4 dsDNA cassettes (Section 3.1.2) were isolated and resuspended in 20µl water (Section 2.1.17). An 18µl volume of each dsDNA cassette was digested with 10 units *Mnl* I at 37°C overnight in 1xNEBuffer 2 (Section 2.1.11) supplemented with 100µg/ml BSA, in a final volume of 40µl. The *Mnl* I was not subsequently heat inactivated, so that the small DNA fragments would not be denatured. A 35µl volume of each digestion reaction was electrophoresed on a 20% non-denaturing polyacrylamide gel, and the released 42 bp N1, 32 bp TGT1, 32 bp TGC1 and 40 bp C1 blocks (Section 3.1.2) were isolated and re-suspended in 15µl water (Section 2.1.17).

A 1µl volume of each N1 and C1 block and a 3µl volume of each TGT1 and TGC1 block was ligated with 400 units T4 DNA ligase (Section 2.1.12). The ligation products were digested with *Not* I / *Sfi* I (Section 2.1.11), separated from the released ends on a 1% LMP agarose gel for 15 minutes (Section 2.1.7) and re-suspended in 15µl water. A 6µl volume of the *Sfi* I / *Not* I digested ligation products was then ligated into 270ng (130 fmol) pUC119- *Sfi* / *Not* -$HIS_6$ (Section 2.1.4) with 400 units T4 DNA ligase (Section 2.1.12). The recombinant DNA constructs were transformed into *E. coli* TOP10F′ cells (Section 2.1.13) and PCR screened for the presence of insert in a $Vent_R$ PCR reaction (Section 2.1.9) using the LacF and LacB primers and the programme: 98°C, 5 min; [94°C, 1 min; 55°C, 2 min; 72°C, 1 min] x 30; 72°C, 5 min. Six inserts were sequenced using the LacB sequencing primer.

### 3.2.5 Assessment of the third strategy for constructing the DNA library

The 58 base C1, 66 base C2, 66 base C3 and 58 base C4 oligonucleotides (Section 3.2.1) were made into dsDNA cassettes (Section 3.2.2) by incubating 10µg of C1 (530pmol), C2 (470pmol), C3 (470pmol) and C4 (530pmol) with 5µg of the Uniprime C (1400pmol) oligonucleotide (Section 3.2.1).

The dsDNA cassettes were not separated from the unreacted oligonucleotides. Instead, a 30µl volume of each Klenow reaction was digested with 5 units *Ear* I at 37°C overnight in 1x NEBuffer 1 (Section 2.1.11), in a final volume of 40µl. The *Ear* I was not subsequently heat inactivated, so that the small DNA fragments would not be denatured. A 35µl volume of each digestion reaction was electrophoresed on a 20% non-denaturing polyacrylamide gel, and the released 40 bp N2, 30 bp TGT2, 30 bp TGC2 and 40 bp C2 blocks (Section 3.1.3) were isolated and re-suspended in 15µl water (Section 2.1.17).

A 4µl volume of each N2, TGT2, TGC2 and C2 block was ligated with 800 units T4 DNA ligase (Section 2.1.12), in a final volume of 20µl. The ligation products were digested with *Not* I / *Sfi* I (Section 2.1.11), separated from the released ends on a 1% LMP agarose gel for 15 minutes (Section 2.1.7) and re-suspended in 10µl water. A 6µl volume of the *Sfi* I / *Not* I digested ligation products was then ligated into 270ng (130 fmol) pUC119- *Sfi* / *Not* -$HIS_6$ (Section 2.1.4.), transformed into *E. coli* TOP10F′ cells and PCR screened for the presence of insert (Section 3.2.4). Five inserts were sequenced using the LacB sequencing primer.

## 3.3 RESULTS AND DISCUSSION

### 3.3.1 Limitations of isolating dsDNA blocks from dsDNA cassettes

The 8 bp NNT1 and 11 bp NNG1 blocks (Section 3.1.1) were originally designed to evaluate the first strategy (Figure 3.1) for constructing the DNA library. However, this strategy was quickly superceded by the second strategy (Figure 3.2) in which the cysteines were encoded at the joint between two ligated blocks, rather than within the blocks themselves (note that random amino acids were encoded at the joint between two ligated blocks in the first strategy). The second strategy had the obvious advantage that less blocks would have to be used to construct the DNA library, since each block only had one variable length spacer. The first strategy was also flawed in the sense that the assembled genes could only encode peptides containing two or more

contiguous cysteines if these cysteines had been introduced by the random amino acids. Note that most conotoxins contain two contiguous cysteines (Section 1.1.2). The only reason why the first strategy is mentioned at all is that the difficulties encountered in isolating the NNT1 and NNG1 blocks illustrate the limitations of forming dsDNA blocks by digesting dsDNA cassettes.

The dsDNA cassettes housing the 8 bp NNT1 and 11 bp NNG1 blocks were successfully formed and isolated (Figure 3.10) and digested with *Mnl* I at 4°C to release the blocks (Figure 3.11)(Section 3.2.3). The *Mnl* I had been incubated at 4°C (even though the recommended temperature was 37°C) and had not been heat inactivated, so that the digestion products would not be denatured. It was then attempted to separate and isolate the 8 bp NNT1 and 11 bp NNG1 blocks from the unwanted 15 bp and 17 bp DNA digestion products, respectively. Both digestion reactions were electrophoresed on a 20% non-denaturing polyacrylamide gel, which has an effective range of separation of 6-100 bp (Sambrook *et al.*, 1989). However, the released 8 bp NNT and 11 bp NNG blocks were not even visible on the polyacrylamide gel. This meant that either not enough of each block was present to show up on the gel, or that the temperatures produced during gel electrophoresis had denatured the blocks. Note that the theoretical melting temperatures of the 8 bp NNT1 and 11 bp NNG1 blocks are approximately 23°C and 32°C, respectively (Sambrook *et al.*, 1989).

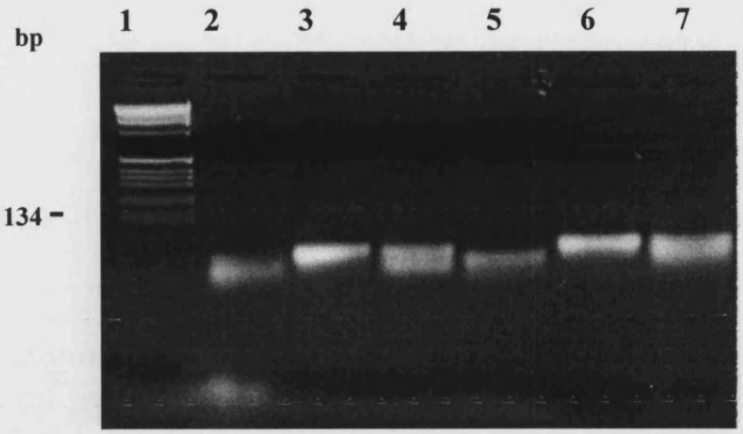



**Figure 3.10** 2% agarose gel displaying the formation of the dsDNA cassettes housing the 8 bp NNT1 and 11 bp NNG1 blocks. Lane 1: 1 kb DNA ladder. Lanes 2 and 4: Oligonucleotide components of the cassette housing the NNT1 block before and after Klenow treatment, respectively. Lane 3: The isolated 40 bp cassette housing the NNT1 block. Lanes 5 and 7: Oligonucleotide components of the cassette housing the NNG1 block before and after Klenow treatment, respectively. Lane 6: The isolated 47 bp cassette housing the NNG1 block.

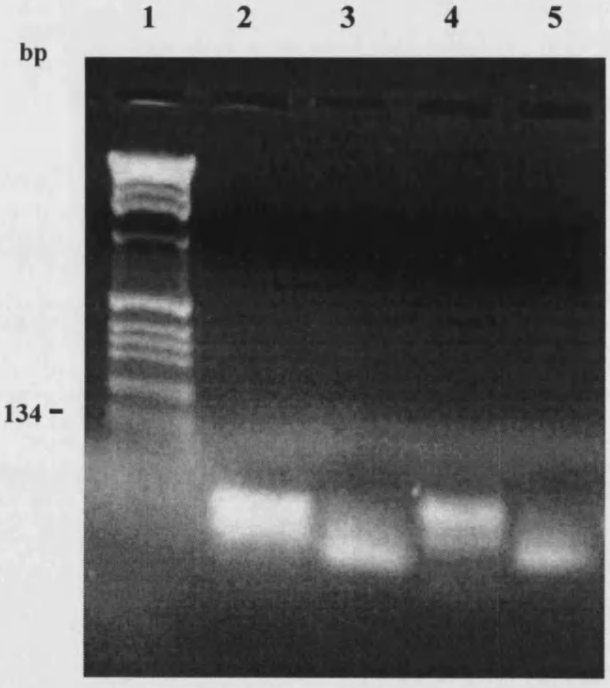



**Figure 3.11** 2% agarose gel displaying the *Mnl* I mediated liberation of the 8 bp NNT1 and 11 bp NNG1 blocks from their parent dsDNA cassettes. Lane 1: 1 kb DNA ladder. Lanes 2 and 3: Cassette housing the NNT1 block before and after *Mnl* I digestion, respectively. Lanes 4 and 5: Cassette housing the NNG1 block before and after *Mnl* I digestion, respectively.

Other methods were investigated as a means of separating and isolating small dsDNA blocks from dsDNA cassette digestion reactions. In one work DNA restriction fragments as small as 7 bp and 11 bp were successfully fractionated using Mono Q anion exchange chromatography (Müller, 1986). A 55μg (19.4pmol) quantity of *Hae* III digested pBR322 (4361 bp) had been applied to a 50 x 5mm Mono Q column in 20mM Tris.HCl (pH 8.2), and eluted with a 20mM Tris.HCl (pH 8.2), 1M NaCl gradient at room temperature (22°C). The two smallest restriction fragments, which were 7 bp and 11 bp in size, were clearly visible as elution peaks on the FPLC trace. However, in order to determine the exact size of dsDNA fragments isolated in this way, the fragments need to be visualised by polyacrylamide gel electrophoresis (Müller, 1986), and the initial problem had been that such small fragments were not visible on these gels.

An alternative method to polyacrylamide gel electrophoresis for determining the size of DNA fragments is mass spectrometry using either electrospray ionization (ESI) or matrix-assisted laser desorption ionization (MALDI)(Cheng *et al.*, 1996)(Andersen *et al.*, 1996)(Fenn *et al.*, 1989). The molecular weight of ssDNA oligonucleotides and dsDNA fragments can be determined from the mass spectrometry spectra (Shaler *et al.*, 1995)(Fenn *et al.*, 1989)(Nordhoff *et al.*, 1993)(Bai *et al.*, 1994)(Wunschel *et al.*, 1998)(Muddiman *et al.*, 1996)(Naito *et al.*, 1995). However, it would be impractical to determine the sizes of the dsDNA blocks using mass spectrometry due to the financial cost (each sample is very expensive and many dsDNA blocks would have to characterised for the construction of each DNA library) coupled with a lack of necessary expertise.

In summary, dsDNA cassettes can be formed from a chemically synthesized oligonucleotide, a suitable oligonucleotide primer and the Klenow fragment; and these cassettes can be digested with restriction endonucleases to liberate building blocks encoding either fixed or random amino acids. However, it is very difficult to isolate small dsDNA blocks (i.e. <12 bp in size) from similarly-sized digestion products by

polyacrylamide gel electrophoresis. The first strategy for constructing our DNA library (Figure 3.1) was never actually tested, as encoding cysteine residues (rather than random amino acids) at the joint between two ligated blocks was deemed a much better design (see above for reasons). Consequently, it was decided to abandon the first strategy and to assess the second strategy (Figure 3.2) for constructing the DNA library instead. Also, it was subsequently determined that building blocks encoding random amino acids could not be used to construct our DNA library anyway (see Chapter 4), thus the first strategy (Figure 3.1) was not viable in any case.

### 3.3.2 Assessment of the second strategy for constructing the DNA library

The second strategy (Figure 3.2) for constructing the DNA library, which uses a one nucleotide 3′ overhang scheme, was evaluated using the N1, TGT1, TGC1 and C1 blocks (Section 3.1.2). The dsDNA cassettes housing these blocks were successfully formed and digested with *Mnl* I to release the blocks (Figure 3.12) which were subsequently isolated from the digestion products (Figure 3.13). The four blocks were then ligated together (Figure 3.13) and cloned into pUC119- *Sfi* I / *Not* I -$HIS_6$ (Section 3.2.4).

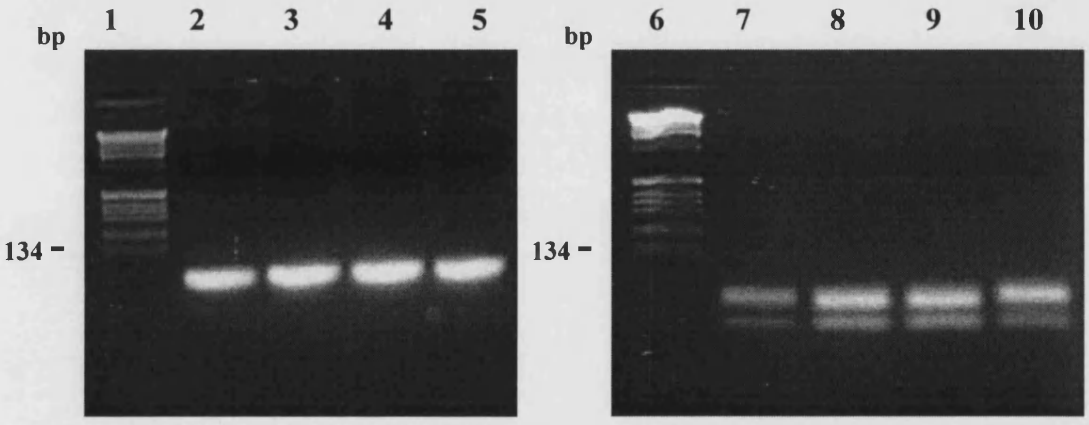



**Figure 3.12** 2% agarose gel displaying the *Mnl* I mediated liberation of the N1, TGT1, TGC1 and C1 blocks from their parent dsDNA cassettes. Lanes 1 and 6: 1 kb DNA ladder. Lane 2: 57 bp cassette housing the N1 block. Lane 3: 62 bp cassette housing the TGT1 block. Lane 4: 62 bp cassette housing the TGC1 block. Lane 5: 55 bp cassette housing the C1 block. Lanes 7-10: The products of lanes 2-5, respectively, after *Mnl* I digestion.

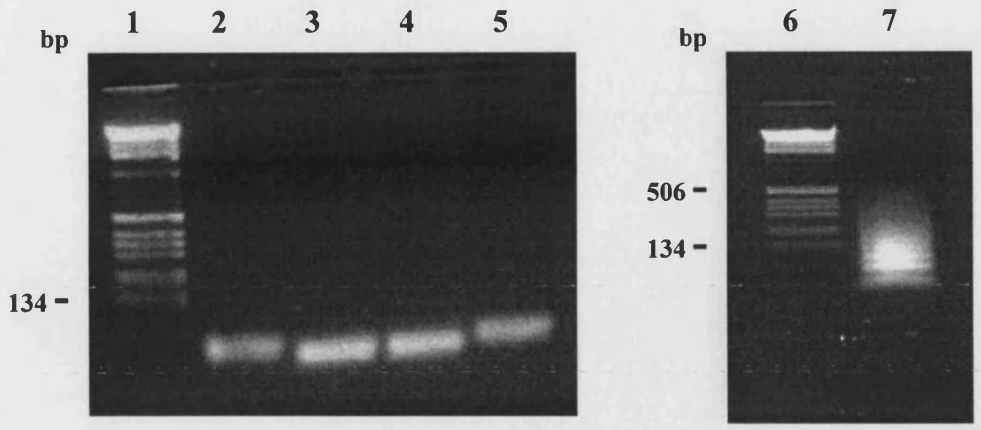



**Figure 3.13** 2% agarose gel displaying the isolated N1, TGT1, TGC1 and C1 blocks and their subsequent ligation. Lanes 1 and 6: 1 kb DNA ladder. Lane 2: 42 bp N1 block. Lane 3: 32 bp TGT1 block. Lane 4: 32 bp TGC1 block. Lane 5: 40 bp C1 block. Lane 7: Ligation of the N1, TGT1, TGC1 and C1 blocks.

A total of 28 transformants were PCR screened. By determining the sizes of the PCR products, the nature of the cloned inserts could be inferred. If the N1, TGT1, TGC1 and C1 blocks ligated together as designed (Figure 3.2) constructs I-IV would be 116 bp, 182 bp, 248 bp and 314 bp in size, respectively. On PCR screening these constructs would produce bands of 314 bp, 380 bp, 446 bp and 512 bp, respectively.

Fourteen transformants (50%) gave bands corresponding to genes encoding an even number of cysteines (Constructs I:9, II:4, III:1) in which the N1, TGT1, TGC1 and C1 blocks had ligated together as intended. No higher constructs (e.g. IV, V, ...) had been assembled. Five inserts were sequenced (Figure 3.14), each sequence confirming the fidelity of the assembly process (Constructs I:2, II:3).

| Insert | | | | | |
|---|---|---|---|---|---|
| **A1** | N1 | TGT1 | TGC1 | TGT1 | C1 |
| **A2** | N1 | TGT1 | TGC1 | TGT1 | C1 |
| **A3** | N1 | TGT1 | C1 | | |
| **A4** | N1 | TGT1 | C1 | | |
| **A5** | N1 | TGT1 | TGC1 | TGT1 | C1 |

**Figure 3.14** Block composition, derived from the nucleotide sequences, of five genes assembled by ligation [N1, TGT1, TGC1, C1].

However, 14 transformants (50%) gave bands corresponding to "illegal" genes in which the N1, TGT1, TGC1 and C1 blocks had ligated together in ways they were not designed to. One insert, which produced a band of 346 bp, was sequenced (Figure 3.15). This insert was an "illegal" gene in which an N1 and TGC1 block had ligated together because both blocks had lost their one nucleotide overhangs, forming a gene which encoded an odd number of cysteines and which was out of frame. This sequence suggested that a large proportion of the "illegal" genes were being formed because the one nucleotide overhangs were being lost from the blocks. It was hypothesized that this was either the result of the one nucleotide overhangs being degraded, or that the *Mnl* I enzyme was digesting the cassettes erroneously. In respect to this point, it may be noted that 80% of DNA fragments produced by a *Mnl* I digestion can be subsequently ligated with T4 DNA ligase (as determined by the manufacturer), implying that 20% of the DNA fragments have been digested incorrectly.

| Insert | | | | |
|---|---|---|---|---|
| **A6** | N1 | TGC1 | TGT1 | C1 |

**Figure 3.15** Block composition, derived from the nucleotide sequence, of an "illegal" gene assembled by ligation [N1, TGT1, TGC1, C1].

This experiment determined that the one nucleotide 3′ overhang scheme utilised by the N, TGT, TGC and C block classes in the second construction strategy (Figure 3.2) is able to assemble genes encoding an even number of cysteines. However, these

genes only constitute approximately 50% of the recombinant DNA constructs, with the other 50% comprising "illegal" genes. To avoid the problems inherent in this one nucleotide overhang scheme, the length of the block overhangs was increased to three nucleotides, in the hope that this would result in a higher proportion of the assembled genes encoding an even number of cysteines.

### 3.3.3 Assessment of the third strategy for constructing the DNA library

The third strategy (Figure 3.3) for constructing the DNA library, which uses a three nucleotide 5′ overhang scheme, was evaluated using the N2, TGT2, TGC2 and C2 blocks (Section 3.1.3). The dsDNA cassettes housing these blocks were successfully formed and digested with *Ear* I to release the blocks (Figure 3.16) which were subsequently isolated from the digestion products (Figure 3.17). The four blocks were then ligated together (Figure 3.17) and cloned into pUC119- *Sfi* I / *Not* I -$HIS_6$ (Section 3.2.5).

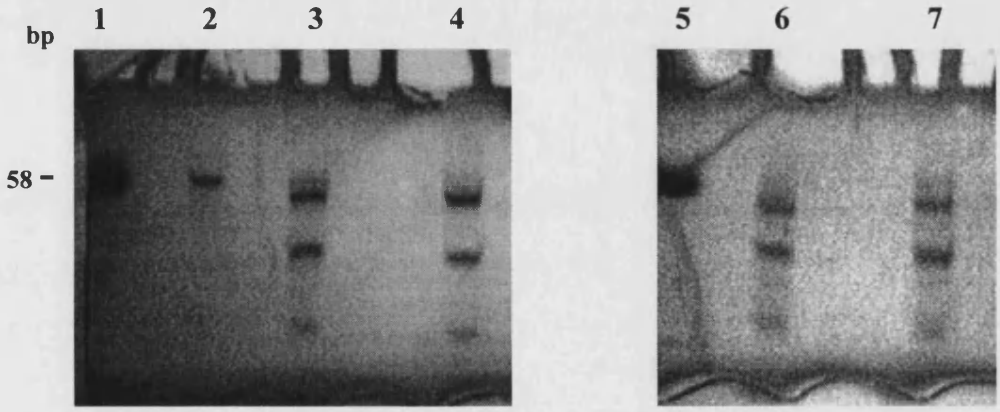



**Figure 3.16** 20% non-denaturing polyacrylamide gel displaying the *Ear* I mediated liberation of the N2, TGT2, TGC2 and C2 blocks from their parent dsDNA cassettes. Lanes 1 and 5: Xylene cyanol FF dye. Lane 2: 58 bp DNA marker. Lane 3: 40 bp N2 block and 15 bp extraneous digestion products. Lane 4: 40 bp C2 block and 15 bp extraneous digestion products. Lane 6: 32 bp TGC2 block and 15 bp extraneous digestion products. Lane 7: 32 bp TGT2 block and 15 bp extraneous digestion products.

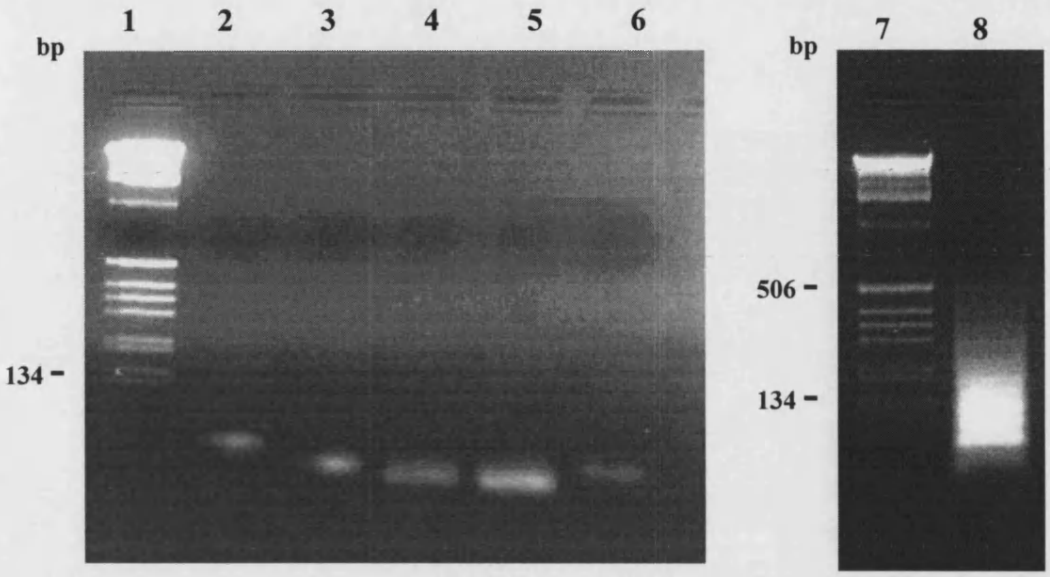



**Figure 3.17** 2% agarose gel displaying the isolated N2, TGT2, TGC2 and C2 blocks and their subsequent ligation. Lanes 1 and 7: 1 kb DNA ladder. Lane 2: 58 bp DNA marker. Lane 3: 40 bp N2 block. Lane 4: 32 bp TGC2 block. Lane 5: 32 bp TGT2 block. Lane 6: 40 bp C2 block. Lane 8: Ligation of the N2, TGT2, TGC2 and C2 blocks.

A total of 23 transformants were PCR screened. By determining the sizes of the PCR products, the nature of the cloned inserts could be inferred. If the N2, TGT2, TGC2 and C2 blocks ligated together as designed (Figure 3.3) constructs I-IV would be 116 bp, 182 bp, 248 bp and 314 bp in size, respectively. On PCR screening, these constructs would produce bands of 314 bp, 380 bp, 446 bp and 512 bp, respectively.

Seventeen transformants (74%) gave bands corresponding to genes encoding an even number of cysteines, in which the N2, TGT2, TGC2 and C2 blocks had ligated together as intended, with a bias towards construct I (I:15, II:1, III:1). No higher constructs (e.g. IV, V, ...) had been assembled. Three inserts were sequenced (Figure 3.18), each sequence confirming the fidelity of the assembly process (Constructs I:2, III:1).

| Insert | | | | | | | |
|---|---|---|---|---|---|---|---|
| **B1** | N2 | TGC2 | C2 | | | | |
| **B2** | N2 | TGC2 | C2 | | | | |
| **B3** | N2 | TGC2 | TGT2 | TGC2 | TGT2 | TGC2 | C2 |

**Figure 3.18** Block composition, derived from the nucleotide sequences, of three genes assembled by ligation [N2, TGT2, TGC2, C2].

However, 6 transformants (26%) gave bands corresponding to genes in which the N2, TGT2, TGC2 and C2 blocks had ligated together "illegally". Two inserts, which produced bands of 281 bp and 413 bp, were sequenced (Figure 3.19). In one "illegal" gene an N2 and TGT2 block had "illegally" ligated together to assemble a gene encoding 5 cysteines which was in frame. In the second "illegal" gene an N2 and C2 block had "illegally" ligated together to assemble a gene encoding 1 cysteine. It is important to note that the overhangs on the blocks which had ligated "illegally" were intact. Both "illegal" genes had been formed because the three nucleotide overhangs on the blocks had violated standard base-pairing. The fact that the three nucleotide overhangs utilised by the N, TGT, TGC and C block classes (Figure 3.3) can disobey standard base-pairing was a disturbing discovery with respect to the strategy, and was investigated in Chapter 4.

| Insert | | | | | | |
|---|---|---|---|---|---|---|
| **B4** | N2 | C2 | | | | |
| **B5** | N2 | TGT2 | TGC2 | TGT2 | TGC2 | C2 |

**Figure 3.19** Block composition, derived from the nucleotide sequences, of two "illegal" genes assembled by ligation [N2, TGT2, TGC2, C2].

This experiment determined that the three nucleotide 5′ overhang scheme utilised by the N, TGT, TGC and C block classes in the third construction strategy (Figure 3.3), is able to assemble genes encoding an even number of cysteines, which constitute approximately 74% of the assembled genes.

The proportion of assembled genes encoding an even number of cysteines is much higher for the N, TGT, TGC and C blocks utilising the 5′ three nucleotide overhang scheme (Figure 3.3), than it is for the blocks utilising the 3′ one nucleotide overhang scheme (Figure 3.2)(Section 3.3.2.). Presumably this is because the three nucleotide overhangs are less prone to degradation. However, the restriction enzyme *Ear* I may well be digesting the dsDNA cassettes more accurately than *Mnl* I did, liberating building blocks which all had three nucleotide overhangs, which may have contributed slightly to the higher proportion of genes encoding an even number of cysteines. With respect to this point, it may be noted that over 95% of DNA fragments produced by an *Ear* I digestion can be subsequently ligated with T4 DNA ligase (as determined by the manufacturer), implying that only 5% of the DNA fragments have been digested incorrectly.

## 3.4 SUMMARY

The basic strategy for constructing a DNA library encoding a population of variable-length peptides which contain an even number of randomly distributed cysteine residues, was successfully determined. The N, TGT, TGC and C classes of sticky-ended dsDNA block, using either the one nucleotide 3′ overhang scheme utilised by the second construction strategy (Figure 3.2) or the three nucleotide 5′ overhang scheme utilised by the third construction strategy (Figure 3.3), are able to assemble genes encoding an even number of cysteines. However, it was decided to use the third construction strategy (Figure 3.3) to construct the DNA library, as the assembled genes encoding an even number of cysteines constituted a higher proportion (74%) of the recombinant DNA constructs.

However, it was also ascertained that the restriction endonuclease digestion of dsDNA cassettes is not a suitable method for forming the dsDNA blocks. This is because a high proportion of the blocks necessary for the construction of the DNA library will be less than 12 bp in size (Chapter 6) and it is very difficult to isolate blocks less than 12 bp in size from the extraneous DNA fragments of these digestion reactions. A

second limitation of this method is that it is very difficult to accurately quantify the amounts of the isolated blocks, indeed the amounts of the blocks added to the gene-assembling ligation reactions in this chapter were estimated visually, after agarose gel electrophoresis. Consequently, a different method of dsDNA block procurement must be determined, a point addressed in Chapter 4.

# CHAPTER 4

# OPTIMISATION OF THE STRATEGY FOR CONSTRUCTING THE DNA LIBRARY

## 4.0 INTRODUCTION

The previous chapter elucidated the core strategy for constructing the DNA library using the N, TGT, TGC and C classes of sticky-ended dsDNA block, which utilise a three nucleotide 5′ overhang scheme (Figure 3.3). This chapter sought to optimise every aspect of the process by which these blocks are to construct the DNA library.

For all arrangements of cysteines to be possible in the DNA library, TGC[ ] and TGT[ ] classes of propagating block (Figure 4.1) which insert between two standard blocks whilst still preserving an even number of cysteine codons, must be used to cater for arrangements involving two adjacent cysteines [CC] and two cysteines separated by a single amino acid [CXC], since TGT and TGC propagating blocks with spacers encoding 0 and 1 amino acids are non-existent and likely to be unstable (Pörschke, 1977), respectively.

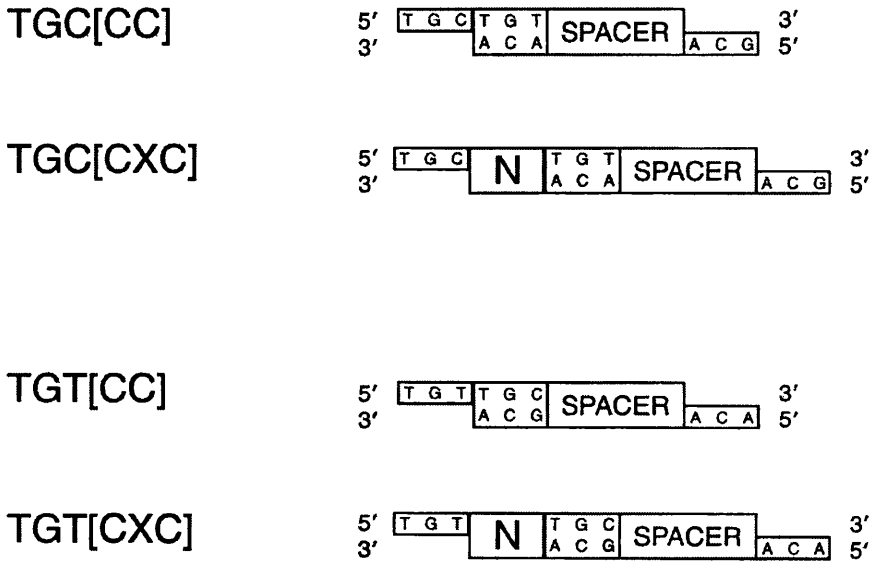



**Figure 4.1** The TGC[ ] and TGT[ ] classes of propagating block. These blocks cater for arrangements involving two adjacent cysteines [CC] and two cysteines separated by a single amino acid [CXC], where X is a single amino acid and N represents 3 base pairs. Note that these blocks encode [CC] and [CXC] at their "N terminus".

The nineteen species of block detailed in this chapter (Section 4.1) representing the N, TGT, TGC, TGT[ ], TGC[ ] and C classes, were prepared by annealing pairs of complementary oligonucleotides together. This approach overcomes many of the problems encountered in the previous chapter. The major advantage is that the blocks are homogenous and require no subsequent purification from either extraneous DNA fragments or enzymic reagents, thus they can be used immediately in a gene-assembling ligation reaction, considerably reducing the set-up time for each experiment. Also, as the concentration of ssDNA oligonucleotides can be accurately determined spectrophotometrically (Section 2.1.5) precise amounts of blocks can be formed. Even very small blocks (i.e. <12 bp in size) can be formed to an accurate concentration and high degree of purity. The only disadvantage is that random amino acids can no longer be encoded in the spacers, as the extreme heterogeneity of a degenerate oligonucleotide mixture would preclude the availability of the complementary oligonucleotide. Consequently, the spacers in the blocks encode fixed amino acids (using the codons most frequently used by *E. coli* [Sambrook *et al.*, 1989]) the significance of which is discussed in Chapter 6.

## 4.1 EXPERIMENTAL DESIGN

### 4.1.1 The TGT, TGC, TGT[ ] and TGC[ ] propagating blocks

**TGC**

```
          C   L
TGC1  5' TGC TTG      3'
      3'     AAC ACA  5'

          C   V   Y
TGC2  5' TGC GTG TAT      3'
      3'     CAC ATA ACA  5'

          C   V   A   Y
TGC3  5' TGC GTG GCA TAT      3'
      3'     CAC CGT ATA ACA  5'

          C   V   A   C   Y
TGC4  5' TGC GTG GCA TGC TAT      3'
      3'     CAC CGT ACG ATA ACA  5'

          C   K   Q   D   S   D
TGC5  5' TGC AAA CAG GAC TCC GAC      3'
      3'     TTT GTC CTG AGG CTG ACA  5'

          C   P   R   I   W   M   E
TGC6  5' TGC CCG CGT ATC TGG ATG GAA     3'
      3'     GGC GCA TAG ACC TAC CTT ACA 5'
```

**TGT**

```
          C   E   F
TGT2  5' TGT GAA TTC      3'
      3'     CTT AAG ACG  5'

          C   G   E   T
TGT3  5' TGT GGT GAA ACT      3'
      3'     CCA CTT TGA ACG  5'

          C   A   T   Y   T
TGT4  5' TGT GCC ACC TAT ACG      3'
      3'     CGG TGG ATA TGC ACG  5'

          C   G   P   N   G   F
TGT5  5' TGT GGT CCG AAC GGT TTC      3'
      3'     CCA GGC TTG CCA AAG ACG  5'

          C   K   Q   S   G   E   M
TGT6  5' TGT AAA CAG TCT GGT GAA ATG     3'
      3'     TTT GTC AGA CCA CTT TAC ACG 5'
```

**TGC[ ]**

```
TGC[CC]5
          C   C   S   T   K   G   E
      5' TGC TGT TCT ACC AAA GGT GAA     3'
      3'     ACA AGA TGG TTT CCA CTT ACG 5'
```

**TGT[ ]**

```
TGT[CC]5
          C   C   S   T   K   G   E
      5' TGT TGC TCT ACC AAA GGT GAA     3'
      3'     ACG AGA TGG TTT CCA CTT ACA 5'
```

**Figure 4.2** Nucleotide sequences and derived amino acid sequences of thirteen species of propagating block representing the TGT, TGC, TGT[ ] and TGC[ ] classes. Note that TGC4 accidently encodes a cysteine within its spacer (this oversight should be ignored).

The TGT and TGC propagating blocks are designated according to the sequence of their 5′ overhang and the number of amino acids encoded by the spacer. The TGT[ ] and TGC[ ] propagating blocks (Figure 4.1) are designated in the same way except that the [ ] symbol preceding the numeral denotes that either CC or CXC (as specified) is encoded at the "N terminus" before the spacer.

**N.B.** In a ligation reaction all the oligonucleotides comprising the blocks must be 5′-phosphorylated, except for the upper strand (the sense strand) of the N terminating block, and the lower strand (the anti-sense strand) of the C terminating block, which must be unphosphorylated to prevent the assembled genes from ligating together.

### 4.1.2 Design of the N-CAPA and C-CAPA terminating blocks

The 27 base (27-mer) PCR A and 27 base (27-mer) PCR B oligonucleotides were designed to anneal to the 12 bp C-CAPA and 12 bp N-CAPA terminating blocks, respectively (Figure 4.3). The last three nucleotides of the PCR A and PCR B primers were designed to anneal to sequences in a TGC block representing a tyrosine codon and an antisense valine codon, respectively.

**Forward primer**

```
                  Hind III
PCR B       5′ CCC AAG CTT CAT GAG GAG GTC TGC GTG            3′

N-CAPA      5′              CAT GAG GAG GTC                   3′
            3′              GTA CTC CTC CAG ACG               5′
```

**Reverse primer (antisense)**

```
C-CAPA      5′              TGT CTC ATG GCA GTC               3′
            3′                  GAG TAC CGT CAG               5′

            3′        ATA ACA GAG TAC CGT CAG AGA TCT CGT    5′    PCR A
                                              Xba I
```

**Figure 4.3** Nucleotide sequences of the PCR A and PCR B primers and the N-CAPA and C-CAPA terminating blocks.

Consequently, any assembled gene in which the N/C-CAPA terminating blocks have each ligated to a TGC block (as they are designed to) will be PCR amplified with the PCR A and PCR B primers. The primers also incorporate *Xba* I and *Hin*d III restriction sites into the PCR products, so that they can be unidirectionally cloned into pCR3 (Figure 4.4).

The advantage of this cloning strategy is that it serves as a means of positive selection. Only those genes in which both terminating blocks have ligated to a TGC block will be PCR amplified and subsequently cloned into pCR3. Any genes not containing both terminating blocks or "illegal" genes in which the terminating blocks have ligated in ways they were not designed to, will not be PCR amplified and subsequently cloned. However, the limitation of this design is that the first and last amino acid encoded in the spacer of each TGC block, has to be fixed as valine and tyrosine, respectively. (Of course it would be possible to redesign the PCR A and PCR B primers so that the last three nucleotides were degenerate, enabling greater variation in the sequences encoded in the TGC spacers).

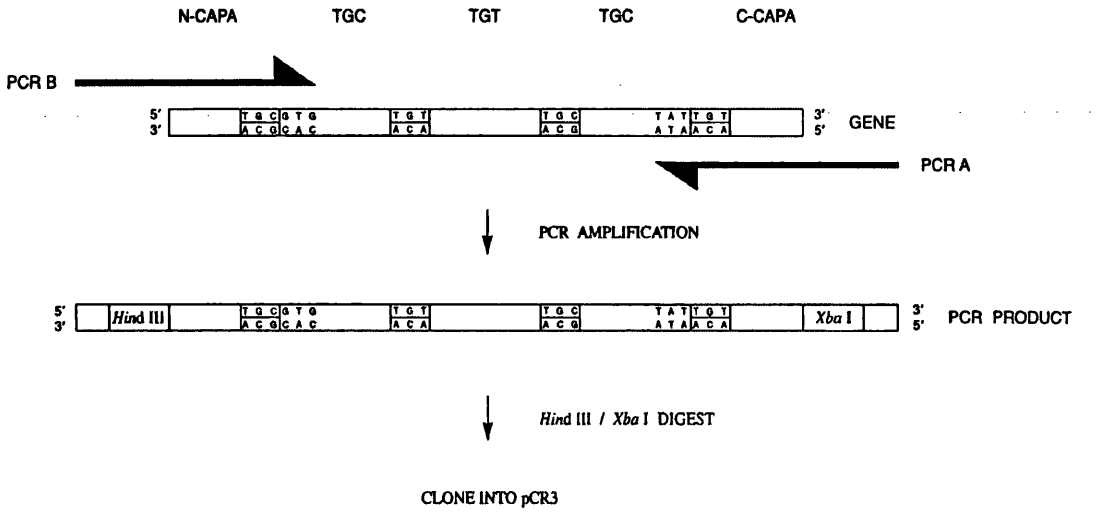



**Figure 4.4** The cloning strategy for genes containing the N/C-CAPA terminating blocks. Genes in which the N/C-CAPA terminating blocks have each ligated to a TGC block are PCR amplified with the PCR A and PCR B primers, which incorporate *Xba* I and *Hind* III restriction sites into the PCR products, so that they can be unidirectionally cloned into pCR3. For example, an NCAPA-TGC4-TGT4-TGC4-CCAPA gene is a type II construct 72 bp in size [(5x12)+(4x3)]. After PCR amplification with the PCR A and PCR B primers, which append 18 bp, this gene would be 90 bp in size.

### 4.1.3 Design of the N-CAPB and C-CAPB terminating blocks

The 27 base (27-mer) PCRL A and 27 base (27-mer) PCRL B oligonucleotides were designed to anneal to the 15 bp N-CAPB and 15 bp C-CAPB terminating blocks, respectively (Figure 4.5). The last three nucleotides of the PCRL A and PCRL B primers are complementary to the nucleotides in the assembled gene encoding the terminal cysteines of the peptide.

**Forward primer**

```
                  Hind III
PCRL A      5' CCC AAG CTT CAT GAG GTC AAC AGG TGC          3'

N-CAPB      5'              CAT GAG GTC AAC AGG              3'
            3'              GTA CTC CAG TTG TCC ACG          5'
```

**Reverse primer (antisense)**

```
C-CAPB      5'        TGT CTC ATG GCA GTC AGC                3'
            3'            GAG TAC CGT CAG TCG                5'

            3'        ACA GAG TAC CGT CAG TCG AGA TCT CGT    5'    PCRL B
                                              Xba I
```

**Figure 4.5** Nucleotide sequence of the PCRL A and PCRL B primers and the N-CAPB and C-CAPB terminating blocks.

Consequently, any assembled gene containing the N/C-CAPB terminating blocks will be PCR amplified with the PCRL A and PCRL B primers. The primers also incorporate *Hind* III and *Xba* I restriction sites into the PCR products, so that they can be unidirectionally cloned into pCR3 (Figure 4.6). The advantage of this cloning strategy is that it imposes no sequence constraints upon the TGC blocks. However, the disadvantage is that the PCRL A and PCRL B primers are able to amplify any assembled gene, "legal" or otherwise, containing the N/C-CAPB terminating blocks.

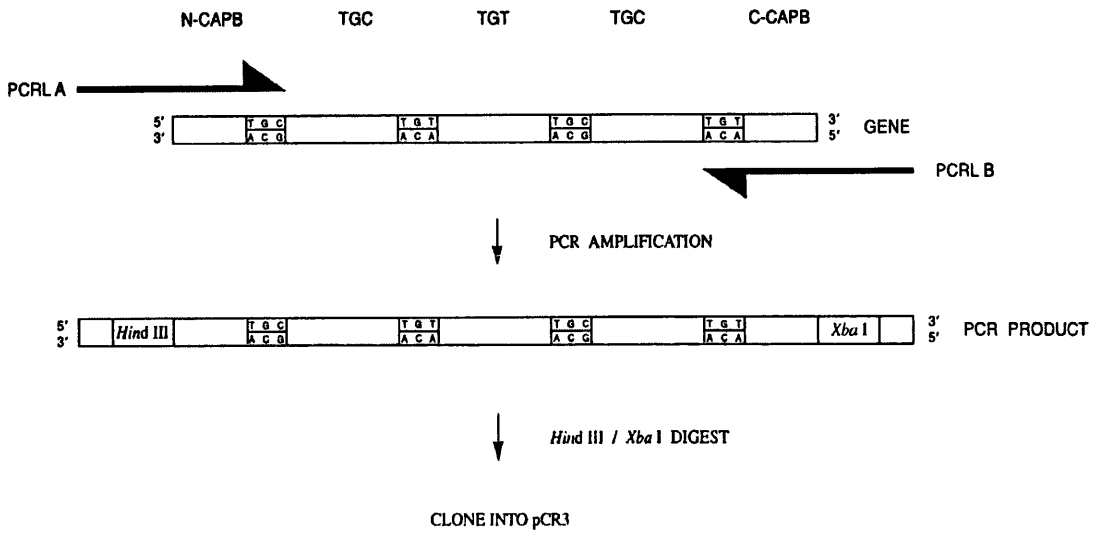



**Figure 4.6** The cloning strategy for genes containing the N/C-CAPB terminating blocks. Genes containing the N/C-CAPB terminating blocks are PCR amplified with the PCRL A and PCRL B primers, which incorporate *Xba* I and *Hind* III restriction sites into the PCR products, so that they can be unidirectionally cloned into pCR3. For example, the NCAPB-TGC4-TGT4-TGC4-CCAPB gene is a type II construct 78 bp in size [(2x15)+(3x12)+(4x3)]. After PCR amplification with the PCRL A and PCRL B primers, which append 18 bp, this gene would be 96 bp in size.

### 4.1.4 Design of the N-CAPC and C-CAPC terminating blocks

The 27 base (27-mer) PCRL A and 27 base (27-mer) PCRL B oligonucleotides were designed to anneal to the 31 bp N-CAPC and 31 bp C-CAPC terminating blocks, respectively (Figure 4.7). The last three nucleotides of the PCRL A and PCRL B primers are complementary to the nucleotides in the assembled gene encoding the terminal cysteines of the peptide.

**Forward primer**

```
                                  Hind III
PCRL A      5'              CCC AAG CTT CAT GAG GTC AAC AGG TGC        3'

N-CAPC      5'    G CTA TAG CCC AAG CTT CAT GAG GTC AAC AGG            3'
            3'    C GAT ATC GGG TTC GAA GTA CTC CAG TTG TCC ACG        5'
```

**Reverse primer (antisense)**

```
C-CAPC      5'        TGT CTC ATG GCA GTC AGC TCT AGA GCA TAC TAT G     3'
            3'            GAG TAC CGT CAG TCG AGA TCT CGT ATG ATA C     5'

            3'        ACA GAG TAC CGT CAG TCG AGA TCT CGT                5'   PCRL B
                                              Xba I
```

**Figure 4.7** Nucleotide sequence of the PCRL A and PCRL B primers and the N-CAPC and C-CAPC terminating blocks.

The disadvantage of the N-CAPB and C-CAPB terminating blocks (Section 4.1.3) is that the assembled genes have to be PCR amplified in order to clone them. To counter this limitation, the N-CAPB and C-CAPB terminating blocks were each elongated by 16 bp, so that the resulting N-CAPC and C-CAPC terminating blocks contain *Hin*d III and *Xba* I restriction sites, respectively.

Consequently, assembled genes containing the N/C-CAPC terminating blocks can either be cloned directly into pCR3 after digestion with *Hin*d III and *Xba* I, or PCR amplified with the PCRL A and PCRL B primers (in the same way as genes containing the N/C-CAPB terminating blocks) and cloned into pCR3 after digestion of the PCR products with *Hin*d III and *Xba* I (Figure 4.8). The advantage of this design is that it offers two cloning strategies.

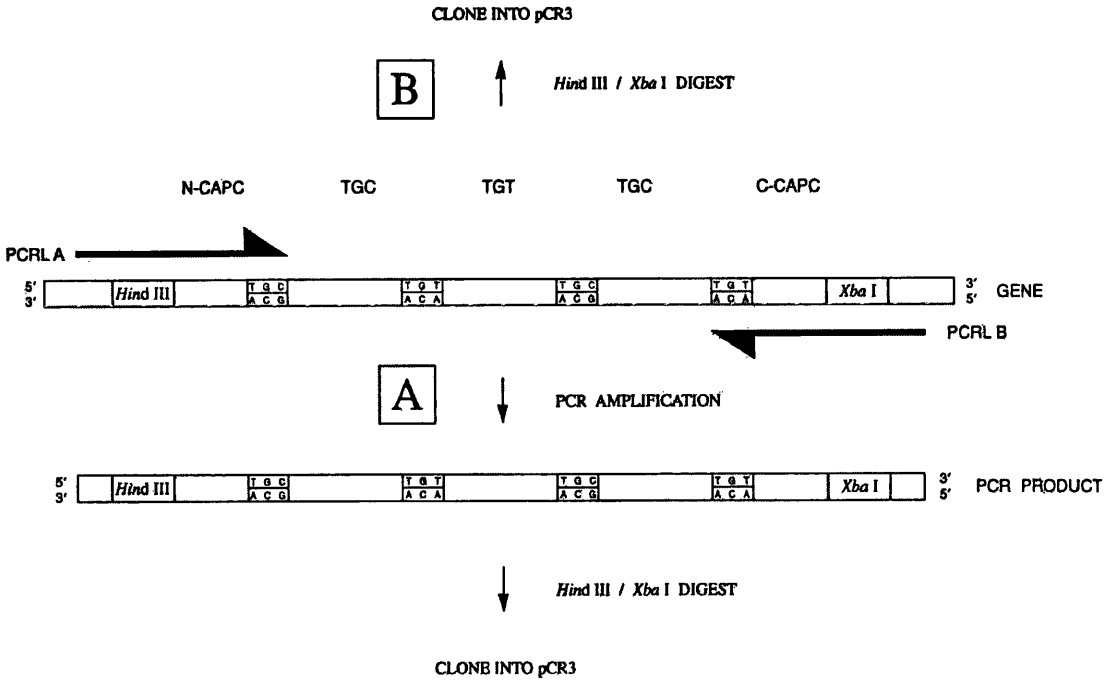



**Figure 4.8** The cloning strategy for genes containing the N/C-CAPC terminating blocks. Genes containing the N/C-CAPC terminating blocks can either be cloned directly into pCR3 after digestion with *Hind* III and *Xba* I [B], or PCR amplified with the PCRL A and PCRL B primers and cloned into pCR3 after digestion of the PCR products with *Hind* III and *Xba* I [A]. For example, the NCAPC-TGC4-TGT4-TGC4-CCAPC gene is a type II construct 110 bp in size [(2x31)+(3x12)+(4x3)]. After PCR amplification with the PCRL A and PCRL B primers, which abolish 18 bp, this gene would be 96 bp in size.

## 4.2 MATERIALS AND METHODS

### 4.2.1 Synthesis of oligonucleotides

All the oligonucleotides in this chapter were synthesized by Perkin-Elmer (Section 2.1.5) although the oligonucleotides comprising the TGC1, TGC4 and TGT4 blocks, the lower strand of the N-CAPA block, and the upper strand of the C-CAPA block, were supplied 5'-phosphorylated. A second set of oligonucleotides comprising the TGC4 and TGT4 blocks were supplied unphosphorylated. When the TGC4 and TGT4 blocks are referred to in the text, the oligonucleotides comprising these blocks were supplied unphosphorylated (unless otherwise stated).

### 4.2.2 Block formation (by temperature-annealing) and subsequent ligation

Complementary oligonucleotides were temperature-annealed to form dsDNA blocks. An equimolar amount (from 1-300pmol, as specified) of the two oligonucleotides constituting a particular block was incubated in 1x T4 DNA ligase buffer (Section 2.1.12) in a final volume of 25µl. The oligonucleotides were initially denatured at 94°C for 1 minute, before being annealed at 37°C for 5 minutes, and cooled to room temperature for 20 minutes. For the ligation reactions, the blocks were mixed together (as specified) and T4 DNA ligase was added to a final concentration of 4 units/µl reaction volume. The ligation reactions were then incubated at 16°C for 2 hours, after which time the T4 DNA ligase was heat inactivated at 65°C for 10 minutes.

### 4.2.3 Block formation (by lyophilisation) and subsequent ligation

An alternative method of annealing complementary oligonucleotides to form dsDNA blocks is lyophilisation. An equimolar amount (from 1-300pmol, as specified) of the two oligonucleotides constituting a particular block was lyophilised (Section 2.1.19). The lyophilised blocks were subsequently re-constituted in 1x T4 DNA ligase buffer (Section 2.1.12) with a 25µl volume of water (pre-chilled to 16°C). For the ligation reactions, the blocks were mixed together (as specified) and T4 DNA ligase was added to a final concentration of 4 units/µl reaction volume. The ligation reactions were then incubated at 16°C for 2 hours, after which time the T4 DNA ligase was heat inactivated at 65°C for 10 minutes.

### 4.2.4 Verification of the experimental strategy

Ligation 1 (Figure 4.9) in a final volume of 100µl, was set up (Section 4.2.2). The oligonucleotides comprising the TGC4 and TGT4 blocks had been supplied phosphorylated.

| Ligation 1 | N-CAPA | TGC4 | TGT4 | C-CAPA |
|---|---|---|---|---|
| pmol | 25 | 150 | 100 | 25 |

**Figure 4.9** Composition of ligation 1 by block species.

A 0.5µl volume of ligation 1 was amplified in a $Vent_R$ PCR reaction using the PCR A and PCR B primers and the programme: [94°C, 1 min; 60°C, 2 min; 72°C, 30 sec] x 25; 72°C, 5 min. The PCR products were purified on a 1% LMP agarose gel (Section 2.1.7), digested with *Hin*d III / *Xba* I (Section 2.1.11), separated from the released ends on a 1.5% LMP agarose gel, and re-suspended in 20µl water. A 6µl volume of the *Hin*d III / *Xba* I digested PCR products was then ligated into 160ng (50 fmol) pCR3 (Sections 2.1.3 and 2.1.12). The recombinant DNA constructs were transformed into *E. coli* TOP10F' cells (Section 2.1.13) and PCR screened for the presence of insert in a $Vent_R$ PCR reaction using the Sp6 and T7 primers and the programme: 98°C, 5 min; [94°C, 1 min; 55°C, 2 min; 72°C, 1 min] x 30; 72°C, 5 min. Ten inserts were sequenced using the pCR3seq primer.

A 0.5µl volume of ligation 1 was amplified in a $Vent_R$ PCR reaction using the PCR A and PCR B primers as described above, except that the number of cycles was increased to 30. The PCR products were then separated on a 12% non-denaturing polyacrylamide gel (Section 2.1.17) and the 60 bp, 90 bp and 120 bp bands were isolated. A 0.5µl volume of each isolated band was then amplified in a $Vent_R$ PCR reaction using the PCR A and PCR B primers as before. The PCR products were purified on a 1.5% LMP agarose gel, digested with *Hin*d III / *Xba* I, separated from the released ends on a 1.5% LMP agarose gel, and re-suspended in 20µl water. A 3µl volume of each *Hin*d III / *Xba* I digested PCR product was then ligated into 80ng (25

fmol) pCR3. The recombinant DNA constructs were transformed into *E. coli* and PCR screened as detailed above. Nine inserts were sequenced using the pCR3seq primer.

### 4.2.5 Can the N, TGT, TGC and C block classes ligate "illegally" ?

Ligations 2-8 (Figure 4.10) each in a final volume of 100µl, were set up (Section 4.2.2). When a particular species of block was to be absent from a reaction, a 25µl volume of 1x T4 DNA ligase buffer was added in its place to maintain the final reaction volume of 100µl. The oligonucleotides comprising the TGC4 and TGT4 blocks had been supplied phosphorylated.

| Ligation | N-CAPA (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPA (pmol) |
|---|---|---|---|---|
| **2** | | 100 | | |
| **3** | | | 100 | |
| **4** | 25 | 150 | 100 | 25 |
| **5** | 25 | 150 | | 25 |
| **6** | 25 | | 100 | 25 |
| **7** | 25 | | | 25 |
| **8** | 50 | | 50 | 50 |

**Figure 4.10** Composition of ligations 2-8 by block species.

A 0.5µl volume of ligations 4-7 was amplified in a *Taq* PCR reaction using the PCR A and PCR B primers and the programme: [94°C, 1 min; 65°C, 1 min; 72°C, 1 min] x 25; 72°C, 5 min.

A 0.5µl volume of ligation 8 was amplified in a $Vent_R$ PCR reaction using the PCR A and PCR B primers and the programme: [94°C, 1 min; 60°C, 2 min; 72°C, 30 sec] x 30; 72°C, 5 min. The PCR products were purified on a 1.5% LMP agarose gel (Section 2.1.7.), digested with *Hind* III / *Xba* I (Section 2.1.11.), separated from the released ends on a 1.5% LMP agarose gel, and finally re-suspended in 15µl water. A 7.5µl volume of the *Hind* III / *Xba* I digested PCR products was then ligated into 80ng (25 fmol) pCR3 and the recombinant DNA constructs were transformed into *E. coli*

and PCR screened (Section 4.2.4). Three inserts were sequenced using the pCR3seq primer.

### 4.2.6 Gene-assembling ligation reactions utilising low amounts of DNA

Ligations 9-10 (Figure 4.11), each in a final volume of 100µl, were set up (Section 4.2.2). The oligonucleotides comprising the TGC4 and TGT4 blocks had been supplied phosphorylated. A 0.5µl volume of each ligation was amplified in a *Taq* PCR reaction using the PCR A and PCR B primers (Section 4.2.5).

| Ligation | N-CAPA (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPA (pmol) |
|---|---|---|---|---|
| **9** | 25 | 150 | 100 | 25 |
| **10** | 1 | 6 | 4 | 1 |

**Figure 4.11** Composition of ligations 9-10 by block species.

### 4.2.7 Incorporation of 6 bp and 9 bp dsDNA blocks

Ligations 11-12 (Figure 4.12) each in a final volume of 100µl, were set up (Section 4.2.2) except that due to the lower melting temperatures of the 6 bp TGC2 and 9 bp TGC3 blocks, the conditions were slightly altered. The blocks in ligation 11 were formed by incubating the complementary oligonucleotide pairs at 94°C for 1 minute and 4°C for 30 minutes, after which time the blocks were ligated together at 4°C for 2 hours. The blocks in ligation 12 were formed by incubating the complementary oligonucleotide pairs at 94°C for 1 minute, 37°C for 15 minutes (room temperature for the TGC3 block), and 16°C for 5 minutes, after which time the blocks were ligated together at 16°C for 2 hours.

| Ligation | N-CAPA (pmol) | TGC2 (pmol) | TGC3 (pmol) | TGT4 (pmol) | C-CAPA (pmol) |
|---|---|---|---|---|---|
| **11** | 12.5 | 50 | | 50 | 12.5 |
| **12** | 12.5 | | 50 | 50 | 12.5 |

**Figure 4.12** Composition of ligations 11-12 by block species.

A 0.5µl volume of ligations 11 and 12 was amplified in a *Taq* PCR reaction using the PCR A and PCR B primers (Section 4.2.5) except that the primer annealing temperature was 60°C. The PCR products were purified on a 1.5% LMP agarose gel (Section 2.1.7), digested with *Hind* III / *Xba* I (Section 2.1.11), separated from the released ends on a 1.5% LMP agarose gel, and finally re-suspended in 30µl water. A 6µl volume of both *Hind* III / *Xba* I digested PCR products was then ligated into 160ng (50 fmol) pCR3 (Sections 2.1.3 and 2.1.12) using 400 units of T4 DNA ligase. The recombinant DNA constructs were transformed into *E. coli* TOP10F′ cells (Section 2.1.13) and PCR screened for the presence of insert in a *Taq* PCR reaction using the Sp6 and T7 primers and the programme: 98°C, 5 min; [94°C, 1 min; 55°C, 1 min; 72°C, 1 min] x 25; 72°C, 5 min. Seven inserts were sequenced using the pCR3seq primer.

### 4.2.8 Optimal means of dsDNA block formation and oligonucleotide 5′-phosphorylation

A 300pmol quantity of each of the two complementary oligonucleotides which comprise the TGT4 block were annealed together by lyophilisation (Section 4.2.3) and temperature-annealing (Section 4.2.2). Both oligonucleotides had been supplied phosphorylated. Both reactions were electrophoresed on a 20% non-denaturing polyacrylamide gel (Section 2.1.17) alongside 300pmol of each individual oligonucleotide in 25µl 1x T4 DNA ligase buffer.

Ligations 13-16 (Figure 4.13), each in a final volume of 100µl, were set up. The only difference between the four ligations was in the method of dsDNA block formation,

and the means by which the 5′-phosphorylated oligonucleotides comprising the propagating blocks had been procured. Ligations 13 and 15 were set up (Section 4.2.3) using blocks formed by lyophilisation, except that prior to being snap-frozen the complementary oligonucleotide pairs were denatured at 94°C for 1 minute. Ligations 14 and 16 were set up (Section 4.2.2) using blocks formed by temperature-annealing. The oligonucleotides comprising the TGC4 and TGT4 blocks in ligations 13-14 had been supplied 5′-phosphorylated, whilst those in ligations 15-16 had been 5′-phosphorylated in the laboratory (Section 2.1.18). A 0.5µl volume of ligations 13-16 was amplified in a *Taq* PCR reaction using the PCR A and PCR B primers (Section 4.2.5) except that the primer annealing temperature was 60°C.

| Ligation | N-CAPA (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPA (pmol) | Blocks | Oligos |
|---|---|---|---|---|---|---|
| **13** | 12.5 | 50 | 50 | 12.5 | Lyophilised | Supplied |
| **14** | 12.5 | 50 | 50 | 12.5 | T-A | Supplied |
| **15** | 12.5 | 50 | 50 | 12.5 | Lyophilised | Laboratory |
| **16** | 12.5 | 50 | 50 | 12.5 | T-A | Laboratory |

**Figure 4.13** Composition of ligations 13-16 by block species, and the means by which the dsDNA blocks had been formed (temperature-annealed [T-A] or lyophilised) and the oligonucleotides (oligos) comprising the propagating blocks had been 5′-phosphorylated (commercially supplied or phosphorylated in the laboratory).

### 4.2.9 The optimal design of the terminating blocks

Ligations 17-18 (Figure 4.14) each in a final volume of 100µl, were set up (Section 4.2.3) except that prior to being snap-frozen the complementary oligonucleotide pairs were denatured at 94°C for 1 minute. Ligation 19 (Figure 4.14), in a final volume of 100µl, was set up (Section 2.1.19). Note that the oligonucleotides comprising the TGC4 and TGT4 blocks in ligation 18 had been supplied phosphorylated, whilst those in ligations 17 and 19 had been phosphorylated in the laboratory (Section 2.1.18).

A 0.5µl volume of ligations 17-19 was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers and the programme: [94°C, 1 min; 60°C, 1 min; 72°C,

1 min] x 25; 72°C, 5 min (note that the primer annealing temperature was raised to 65°C for ligation 19, and that the PCR A and PCR B primers were used to amplify ligation 17).

| Ligation 17 | N-CAPA | TGC4 | TGT4 | C-CAPA |
|---|---|---|---|---|
| pmol | 2 | 10 | 10 | 2 |

| Ligation 18 | N-CAPB | TGC4 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 2 | 10 | 10 | 2 |

| Ligation 19 | N-CAPC | TGC4 | TGT4 | C-CAPC |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

**Figure 4.14** Composition of ligations 17-19 by block species.

### 4.2.10 Physical optimisation of the gene-assembling ligation reactions

#### Determination of ligation temperature

Ligations 20-23 (Figure 4.15), each in a final volume of 100µl, were set up (Section 4.2.3) except that prior to being snap-frozen the complementary oligonucleotide pairs for ligation 20 were denatured at 94°C for 1 minute, and the lyophilised blocks in ligation 23 were re-suspended in water pre-chilled to 4°C and were subsequently ligated together at 4°C. A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers and the programme: [94°C, 1 min; 65°C, 1 min; 72°C, 1 min] x 25; 72°C, 5 min.

| Ligation | N-CAPB (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPB (pmol) | Oligonucleotides Denatured Prior to Lyophilisation? |
|---|---|---|---|---|---|
| **20** | 2.5 | 10 | 6 | 2.5 | Yes |
| **21** | 2.5 | 10 | 6 | 2.5 | No |

| Ligation | N-CAPB (pmol) | TGC2 (pmol) | TGT6 (pmol) | C-CAPB (pmol) | Ligation Temperature |
|---|---|---|---|---|---|
| **22** | 2.5 | 10 | 6 | 2.5 | 16°C |
| **23** | 2.5 | 10 | 6 | 2.5 | 4°C |

**Figure 4.15** Composition of ligations 20-23 by block species. Also detailed are the ligation temperatures (for ligations 22-23) and whether the complementary oligonucleotides comprising the blocks were denatured at 94°C for 1 minute prior to lyophilisation (for ligations 20-21).

### Optimisation of the N:TGC:TGT:C molar ratio

Ligations 24-37 (Figure 4.16) each in a final volume of 100µl, were set up (Section 4.2.3). A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers.

| Ligation | N-CAPB (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPB (pmol) |
|---|---|---|---|---|
| **24** | 2.5 | 2.5 | 2.5 | 2.5 |
| **25** | 2.5 | 5 | 5 | 2.5 |
| **26** | 2.5 | 10 | 10 | 2.5 |
| **27** | 2.5 | 10 | 6 | 2.5 |
| **28** | 2.5 | 6 | 10 | 2.5 |
| **29** | 2.5 | 10 | 10 | 2.5 |
| **30** | 2.5 | 10 | 7.5 | 2.5 |
| **31** | 2.5 | 10 | 6.66 | 2.5 |
| **32** | 2.5 | 10 | 6 | 2.5 |
| **33** | 2.5 | 10 | 5 | 2.5 |
| **34** | 1 | 10 | 6 | 1 |
| **35** | 2 | 10 | 6 | 2 |
| **36** | 2.5 | 10 | 6 | 2.5 |
| **37** | 4 | 10 | 6 | 4 |

**Figure 4.16** Composition of ligations 24-37 by block species.

**Efficacy of two-stage ligation reactions**

Ligations 38-41 (Figure 4.17) each in a final volume of 100µl, were set up (Section 4.2.3) except that in ligations 38, 39, 40 and 41 the propagating blocks (TGC4 and TGT4) were ligated together at 16°C with 200 units T4 DNA ligase (4 units/µl reaction volume) for a period of 0, 30, 60 and 120 minutes, respectively, in a 50µl primary ligation. The N-CAPB and C-CAPB terminating blocks were then added, with a further 200 units T4 DNA ligase (to preserve the 4 units/µl reaction volume concentration) giving a final reaction volume of 100µl, and the ligations were incubated at 16°C for a further 120 minutes. Note that before addition, the resuspended terminating blocks had been kept separate at 16°C. A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers.

| Ligation | N-CAPB (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPB (pmol) | Primary Ligation (minutes) | Secondary Ligation (minutes) |
|---|---|---|---|---|---|---|
| **38** | 2.5 | 10 | 6 | 2.5 | 0 | 120 |
| **39** | 2.5 | 10 | 6 | 2.5 | 30 | 120 |
| **40** | 2.5 | 10 | 6 | 2.5 | 60 | 120 |
| **41** | 2.5 | 10 | 6 | 2.5 | 120 | 120 |

**Figure 4.17** Composition of ligations 38-41 by block species, and the durations of the primary and secondary ligation reactions.

**Non-preferential incorporation of different sized blocks**

Ligations 42-44 (Figure 4.18) each in a final volume of 100µl, were set up (Section 2.1.19). Note that the five blocks in ligation 43 were formed and re-suspended in 20µl volumes so as to maintain the final reaction volume of 100µl. A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers.

| Ligation | N-CAPB (pmol) | TGC2 (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPB (pmol) |
|---|---|---|---|---|---|
| 42 | 1 | | 10 | 6 | 1 |
| 43 | 1 | 5 | 5 | 6 | 1 |
| 44 | 1 | 10 | | 6 | 1 |

**Figure 4.18** Composition of ligations 42-44 by block species.

### Optimum temperature and duration of ligation reactions

Ligations 45-48 (Figure 4.19) each in a final volume of 100µl, were set up (Section 2.1.19) except that ligations 47-48 used a ligation temperature of 16°C, and the duration of the secondary ligation reaction in ligations 46 and 48 was 14 hours (overnight). Oligonucleotides that had been supplied phosphorylated were used to form the TGC4 and TGT4 blocks, so that large amounts of these blocks could be formed by lyophilisation (Section 2.1.19).

| Ligation | N-CAPC (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPC (pmol) | Ligation Temperature | Primary Ligation (minutes) | Secondary Ligation (minutes) |
|---|---|---|---|---|---|---|---|
| 45 | 6 | 60 | 36 | 6 | 4°C | 60 | 120 |
| 46 | 6 | 60 | 36 | 6 | 4°C | 60 | Overnight |
| 47 | 6 | 60 | 36 | 6 | 16°C | 60 | 120 |
| 48 | 6 | 60 | 36 | 6 | 16°C | 60 | Overnight |

**Figure 4.19** Composition of ligations 45-48 by block species. The temperature and duration of the primary and secondary ligation is also detailed for each reaction.

### An excess of terminating blocks in the secondary ligation

Ligations 49-51 (Figure 4.20) each in a final volume of 100µl, were set up (Section 2.1.19). A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers.

| Ligation | N-CAPC (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPC (pmol) |
|---|---|---|---|---|
| 49 | 1 | 10 | 6 | 1 |
| 50 | 4 | 10 | 6 | 4 |
| 51 | 10 | 10 | 6 | 10 |

**Figure 4.20** Composition of ligations 49-51 by block species.

### 4.2.11 Non-incorporation of 3 bp dsDNA blocks

Ligations 52-56 (Figure 4.21) each had a final volume of 100µl. Ligations 55-56 were set up (Section 2.1.19) using a ligation temperature of 4°C. When a species of block was to be absent from a ligation reaction, a 25µl volume of 1x T4 DNA ligase buffer was added in its place to maintain the final volume of 100µl. Ligations 52-53 were set up (Section 4.2.2) except that the blocks were formed and subsequently ligated together in PCR tubes in a thermal cycler (Section 2.1.9). The complementary oligonucleotide pairs were denatured at 94°C for 1 minute, and annealed at 0.5°C for 15 minutes to form the blocks, which were then ligated together (Section 2.1.19) at a ligation temperature of 0.5°C (instead of 4°C). Ligation 54 was set up in the same way as ligations 52-53 except that the temperature was raised from 0.5°C to 2°C. A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers (Section 4.2.10).

| Ligation | N-CAPB (pmol) | TGC1 (pmol) | TGC4 (pmol) | TGT4 (pmol) | C-CAPB (pmol) | Ligation Temperature |
|---|---|---|---|---|---|---|
| 52 | 1 | | 10 | 6 | 1 | 0.5°C |
| 53 | 1 | 10 | | 6 | 1 | 0.5°C |
| 54 | 1 | 10 | | 6 | 1 | 2°C |
| 55 | 1 | 10 | | 6 | 1 | 4°C |
| 56 | 1 | | | 6 | 1 | 4°C |

**Figure 4.21** Composition of ligations 52-56 by block species, and the ligation temperatures.

### 4.2.12 Incorporation of the TGT2-6 and TGC2-6 propagating blocks

Ligations 57-66 (Figure 4.22) each in a final volume of 100µl, were set up (Section 2.1.19). A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers (Section 4.2.10). Ligations 64-66 used the N/C-CAPLC terminating blocks (Section 4.1.4). Constructs containing these terminating blocks are PCR amplified with the PCRL A and PCRL B primers in the same way as constructs containing the N/C-CAPB terminating blocks, thus the PCR products of ligations 64-66 are directly comparable in size to the PCR products of ligations 57-63.

| Ligation 57 | N-CAPB | TGC3 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 58 | N-CAPB | TGC4 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 59 | N-CAPB | TGC5 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 60 | N-CAPB | TGC6 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 61 | N-CAPB | TGC4 | TGT3 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 62 | N-CAPB | TGC4 | TGT5 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 63 | N-CAPB | TGC4 | TGT6 | C-CAPB |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 64 | N-CAPC | TGC4 | TGT4 | C-CAPC |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 65 | N-CAPC | TGC2 | TGT4 | C-CAPC |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

| Ligation 66 | N-CAPC | TGC4 | TGT2 | C-CAPC |
|---|---|---|---|---|
| pmol | 1 | 10 | 6 | 1 |

**Figure 4.22** Composition of ligations 57-66 by block species.

### 4.2.13 The TGC[ ] and TGT[ ] classes of block

Ligations 67-72 (Figure 4.23) were set up (Section 2.1.19). The final volumes of the reactions were 100µl (ligations 67-69), 125µl (ligations 70-71) and 150µl (ligation 72). As the TGC[ ] and TGT[ ] classes of block are essentially TGT and TGC blocks ligated together (e.g. 6pmol TGC[CC]5 is the equivalent to 6pmol TGC0 ligated to 6pmol TGT5, although not the species of TGT5 block detailed in Figure 4.2), it is possible to represent the composition of ligation reactions containing these blocks in terms of only the N, TGT, TGC and C classes of block. The optimum N:TGC:TGT:C molar ratio of 1:10:6:1 (Section 4.3.7) was preserved in all the ligation reactions. A 0.5µl volume of each ligation reaction was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers (Section 4.2.10).

| Ligation | N-CAPC (pmol) | TGC3 (pmol) | TGC4 (pmol) | TGT4 (pmol) | TGC[CC]5 (pmol) | TGT[CC]5 (pmol) | C-CAPC (pmol) |
|---|---|---|---|---|---|---|---|
| **67** | 1 | 4 | | | 6 | | 1 |
| **68** | 1 | 4 | | | | 6 | 1 |
| **69** | 1 | | 10 | 6 | | | 1 |
| **70** | 1 | | 7 | 3 | 3 | | 1 |
| **71** | 1 | | 7 | 3 | | 3 | 1 |
| **72** | 1 | | 6 | 2 | 2 | 2 | 1 |

**Figure 4.23** Composition of ligations 67-72 by block species.

The *Taq* PCR reaction of ligation 72 was alcohol precipitated (Section 2.1.8), digested with *Hin*d III / *Xba* I (Section 2.1.11), separated from the released ends on a 1.5% LMP agarose gel (Section 2.1.7), and re-suspended in 20µl water. A 5.5µl volume of the *Hin*d III / *Xba* I digested PCR products was then ligated into 240ng (75 fmol) pCR3 (Sections 2.1.3 and 2.1.12). The recombinant DNA constructs were transformed into *E. coli* TOP10F′ cells (Section 2.1.13) and PCR screened for the presence of insert in a *Taq* PCR reaction using the Sp6 and T7 primers and the programme: 94°C, 5 min [94°C, 1 min; 55°C, 1 min; 72°C, 1 min] x 30; 72°C, 5 min. Five inserts were sequenced using the pCR3seq primer.

### 4.2.14 Elucidation of the "double-band" phenomenon

Ligation 73 (Figure 4.24), in a final volume of 100µl, was set up (Section 4.2.3) except that prior to being snap-frozen the complementary oligonucleotide pairs were denatured at 94°C for 1 minute. The oligonucleotides comprising the TGC4 and TGT4 blocks had been supplied phosphorylated. A 0.5µl volume of ligation 73 was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers and the programme: [94°C, 1 min; 60°C, 1 min; 72°C, 30 sec] x 25; 72°C, 5 min.

| Ligation 73 | N-CAPB | TGC4 | TGT4 | C-CAPB |
|---|---|---|---|---|
| pmol | 2 | 10 | 6 | 2 |

**Figure 4.24** Composition of ligation 73 by block species.

Four cloned inserts, representing the first four types of construct (I-IV) assembled by ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA) and which had been verified by sequencing (Section 4.3.1) were PCR amplified. A 0.5µl volume of SNAP miniprep (at a 1:100 dilution) (Section 2.1.16) of each recombinant DNA construct, was amplified in a *Taq* PCR reaction using the PCR A and PCR B primers and the programme: [94°C, 1 min; 55°C, 1 min; 72°C, 1 min] x 30; 72°C, 5 min.

Four combinations of the PCR products of constructs I-IV were then mixed together in duplicate (I, II, III and IV; III and IV; II and III; I and II). Each reaction had a final volume of 12µl, and 3µl volumes of the PCR reactions were added as detailed. One set of reactions was heated at 94°C for 1 minute, 37°C for 5 minutes, followed by a 10 minute incubation at room temperature, in order to denature and re-anneal the PCR products. The other identical set of reactions was left untreated at room temperature. The four denaturing/re-annealing reactions were visualised, alongside their untreated pairs and the four original PCR reactions on a 3% agarose gel (Section 2.1.6).

## 4.3 RESULTS AND DISCUSSION

### 4.3.1 Verification of the experimental strategy

Amplification of ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA) (Section 4.2.4) with primers specific for the N/C-CAPA blocks gave a ladder of bands diagnostic of correct gene formation (Figure 4.25), each band corresponding in size to a gene encoding an even number of cysteine residues, with the four smallest representing the first four types of construct (I:60bp, II:90bp, III:120bp, IV:150bp). Note that the negative PCR controls produced a band of approximately 45 bp, which was assumed to be a primer-dimer.

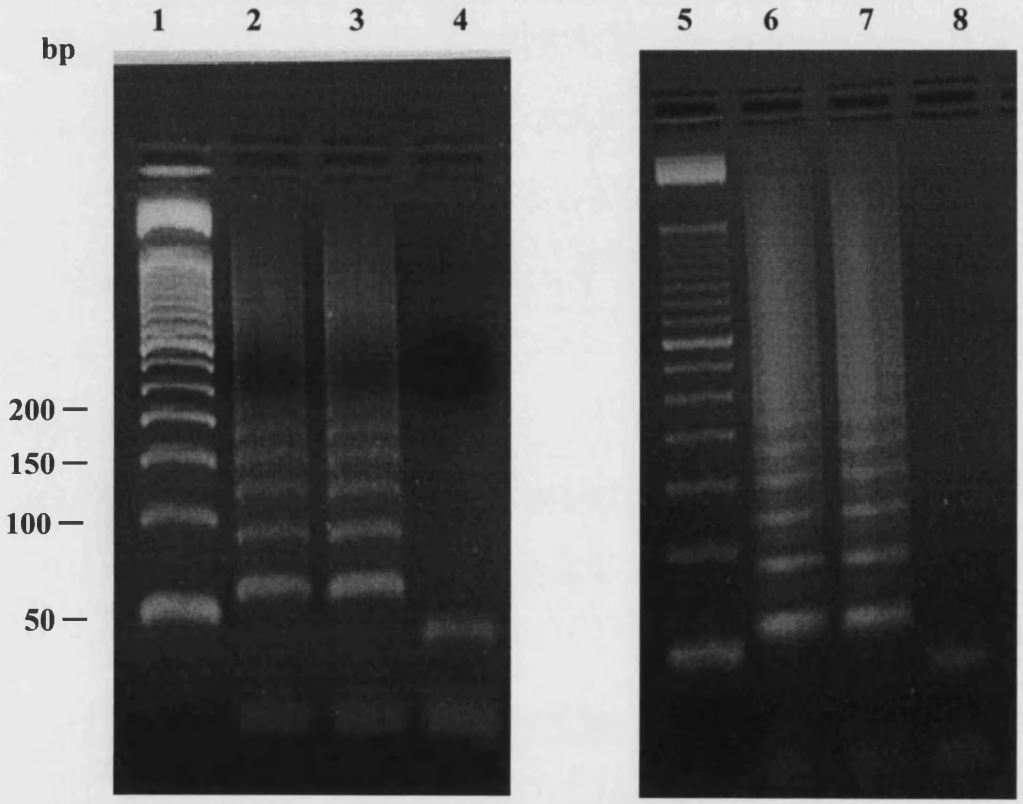



**Figure 4.25** Experimental verification of the ability to form dsDNA blocks by annealing complementary oligonucleotides and of the N, TGT, TGC and C block classes to ligate together as intended. Each band corresponds to a gene assembled by ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA) encoding an even number of cysteines, with the four smallest representing constructs I-IV encoding 2, 4, 6, and 8 cysteines, respectively. Lane 1: 50 bp DNA ladder. Lanes 2 and 3: PCR amplification (25 cycles) of ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA). Lane 4: Negative PCR control. Lane 5: 50 bp DNA ladder. Lanes 6 and 7: PCR amplification (30 cycles) of ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA). Lane 8: Negative PCR control.

In order to unequivocally determine the nature of the assembled genes the PCR products were cloned into pCR3 and ten inserts were sequenced (Figure 4.26). Six inserts were genes encoding an even number of cysteines (Constructs I:2, II:2, IV:2) in which the blocks had ligated together as intended. However, two inserts were "illegal" genes, in which the blocks had ligated together in ways they were not designed to (note that the block overhangs were intact). Three inserts were "double-constructs" in which two genes had ligated together because a *Xba* I undigested C-CAPA block in one assembled gene had ligated to a *Hin*d III digested N-CAPA block in a second gene.

| Insert | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **A1** | N-CAPA | TGC4 | TGT4 | TGC4 | C-CAPA | | | | |
| **A2** | N-CAPA | TGC4 | C-CAPA | | | | | | |
| **A3** | N-CAPA | TGC4 | C-CAPA | N-CAPA | TGC4 | C-CAPA | | | |
| **A4** | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA |
| **A5** | N-CAPA | TGC4 | C-CAPA | | | | | | |
| **A6** | N-CAPA | TGC4 | C-CAPA | N-CAPA | TGC4 | C-CAPA | | | |
| **A7** | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA |
| **A8** | N-CAPA | TGC4 | C-CAPA | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 |
| | TGT4 | TGC4 | TGC4 | C-CAPA | | | | | |
| **A9** | N-CAPA | TGC4 | TGT4 | TGC4 | C-CAPA | | | | |
| **A10** | N-CAPA | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA | | | |

**Figure 4.26** Block composition, derived from the nucleotide sequences, of ten genes assembled by ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA). Note that inserts A3, A6 and A8 are "double-constructs" formed due to incomplete *Xba* I digestion of the assembled genes, and that inserts A8 and A10 are "illegal" genes.

In a second experiment, the 60 bp, 90 bp and 120 bp bands were individually isolated from the PCR products and cloned into pCR3. Nine inserts (3 from each band) were sequenced (Figure 4.27). All nine inserts were "legal" genes, representing the first three types of construct (I:3, II:3, III:3). This confirmed that the 60 bp, 90 bp and 120 bp bands were constructs I, II and III, encoding 2, 4, and 6 cysteines, respectively.

| Insert | | | | | | | |
|---|---|---|---|---|---|---|---|
| **B1** | N-CAPA | TGC4 | C-CAPA | | | | |
| **B2** | N-CAPA | TGC4 | C-CAPA | | | | |
| **B3** | N-CAPA | TGC4 | C-CAPA | | | | |
| **B4** | N-CAPA | TGC4 | TGT4 | TGC4 | C-CAPA | | |
| **B5** | N-CAPA | TGC4 | TGT4 | TGC4 | C-CAPA | | |
| **B6** | N-CAPA | TGC4 | TGT4 | TGC4 | C-CAPA | | |
| **B7** | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA |
| **B8** | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA |
| **B9** | N-CAPA | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPA |

**Figure 4.27** Block composition, derived from the nucleotide sequences, of a further nine genes assembled by ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA).

This experiment served as a keystone in determining the success of preparing dsDNA blocks by annealing complementary oligonucleotides, the efficacy of the cloning strategy (Section 4.1.2) and the fidelity of the assembly process (Figure 3.3). However, it had also shown that the N, TGT, TGC and C block classes are occasionally able to ligate together "illegally", a phenomenon first seen in Chapter 3.

### 4.3.2 The N, TGT, TGC and C block classes can ligate "illegally"

The N, TGT, TGC and C classes of block had been specifically designed to assemble genes encoding an even number of cysteines (Figure 3.3). The inability of the three nucleotide overhangs utilised by these blocks to disobey standard base-pairing was central to this design. The objective of this experiment (Section 4.2.5) was to investigate the ability of the four block classes to ligate together "illegally".

#### "Illegal" ligation of the N and C terminating blocks

Theoretically, ligation 7 (N-CAPA, C-CAPA) should not assemble any genes at all. However, a single DNA construct was formed probably by the two terminating blocks "illegally" ligating together (Figure 4.28, lane 7). No PCR amplification was possible

(lane 12) probably because the PCR A and PCR B primers could not anneal correctly (Section 4.1.2). However, a very faint band was produced which was probably caused by a contaminant in the PCR reaction.

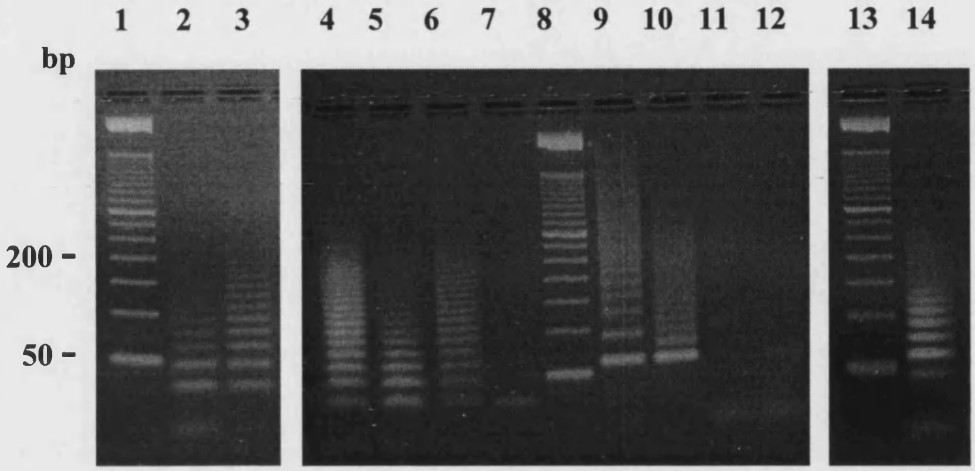



**Figure 4.28** "Illegal" ligation capabilities of the N, TGT, TGC and C block classes. Lane 1: 50 bp DNA ladder. Lane 2: Ligation 2 (TGC4) demonstrating TGC self-polymerisation. Lane 3: Ligation 3 (TGT4) demonstrating TGT self-polymerisation. Lane 4: Ligation 4 (N-CAPA, TGC4, TGT4, C-CAPA) demonstrating correct gene assembly. Lane 5: Ligation 5 (N-CAPA, TGC4, C-CAPA) demonstrating "illegal" gene assembly. Lane 6: Ligation 6 (N-CAPA, TGT4, C-CAPA) demonstrating "illegal" gene assembly. Lane 7: Ligation 7 (N-CAPA, C-CAPA) demonstrating "illegal" gene assembly. Lane 8: 50 bp DNA ladder. Lanes 9-12: The products of lanes 4-7, respectively, after PCR amplification with the PCR A/B primers. Lane 13: 50 bp DNA ladder. Lane 14: PCR amplification of ligation 8 (N-CAPA, TGT4, C-CAPA) demonstrating "illegal" gene assembly.

### TGT self-polymerisation

The ability of the TGT class of block to "illegally" self-polymerise is clearly illustrated by ligation 3 (TGT4) (lane 3). Theoretically, ligation 6 (N-CAPA, TGT4, C-CAPA) should not assemble any genes. However, the products (lane 6) appeared as a ladder of bands. These genes had probably been formed by the N and C terminating blocks "illegally" ligating to the "illegally" self-polymerising TGT blocks. No PCR amplification of these "illegal" genes (lane 11) was possible, because the PCR A and PCR B primers could not anneal correctly (Section 4.1.2).

**TGC self-polymerisation**

The TGC class of block was also able to "illegally" self-polymerise, as illustrated by ligation 2 (TGC4) (lane 2), although the extent of large construct formation was less than for the TGT4 block (lane 3) suggesting that the TGT class of block is more susceptible to self-polymerisation. Theoretically, ligation 5 (N-CAPA, TGC4, C-CAPA) should only assemble a single type I construct encoding 2 cysteines. However, the products appeared as a ladder of bands (lane 5). PCR amplification of these "illegal" genes with primers specific for the N/C-CAPA blocks, gave a ladder of bands corresponding in size to genes encoding 1, 2, ... cysteines (lane 10). These genes had probably been formed by the N and C terminating blocks ligating to the "illegally" self-polymerising TGC blocks.

**The N, TGT, TGC and C block classes preferentially ligate together as designed**

The products of ligation 4 (N-CAPA, TGC4, TGT4, C-CAPA) appeared as a ladder of bands increasing in increments of 15 bp (lane 4). Since each of the four species of block was 12 bp in size (with 3nt overhangs), each 15 bp increase probably represented the addition of a single block to the growing gene. The intermediate bands between the different types of assembled construct (I:42bp, II:72bp, III:102bp, IV:132bp) were hypothesised to be partial constructs which had not been capped at both ends with the terminating blocks. Amplification with primers specific for the N/C-CAPA blocks gave a ladder of bands (lane 9) diagnostic of correct gene assembly (Construct I:60bp, II:90bp, III:120bp, IV:150bp) proving that the four classes of block had ligated together as designed.

**"Illegal" genes are formed by the overhangs disobeying standard base-pairing**

Ligation 8 (N-CAPA, TGT4, C-CAPA) was amplified in a $Vent_R$ PCR reaction with primers specific for the N/C-CAPA blocks, the products appearing as a ladder of

bands corresponding in size to genes encoding 1, 2, ... cysteines (lane 14). These PCR products were cloned into pCR3. Of 15 screened transformants, three inserts were sequenced (Figure 4.29). All three inserts were "illegal" genes, confirming that the N and C terminating blocks can "illegally" ligate together, and that these terminating blocks can "illegally" ligate to the "illegally" self-polymerising TGT blocks. All these "illegal" genes had been assembled because the overhangs on the blocks had disobeyed standard base-pairing (none of the overhangs had been lost from the blocks). $Vent_R$ DNA polymerase, in contrast to *Taq* DNA polymerase, contains an integral 3′→5′ proof-reading exonuclease activity (Mattila *et al.*, 1991)(Kong *et al.*, 1993) which would have removed the mismatched nucleotides at the 3′ ends of the PCR A and PCR B primers once they had annealed to the "illegal" genes. This explains why the PCR A and PCR B primers were able to amplify the "illegal" genes which they did not properly anneal to (Section 4.1.2).

| Insert | | | | | |
|---|---|---|---|---|---|
| C1 | N-CAPA | TGT4 | C-CAPA | | |
| C2 | N-CAPA | TGT4 | TGT4 | TGT4 | C-CAPA |
| C3 | N-CAPA | C-CAPA | | | |

**Figure 4.29** Block composition, derived from the nucleotide sequences, of three "illegal" genes assembled by ligation 8 (N-CAPA, TGT4, C-CAPA).

**Summary**

This experiment determined that the three nucleotide overhangs utilised by the N, TGT, TGC and C block classes (Figure 3.3) are able to disobey standard base-pairing under certain conditions. This enables the block classes to ligate together "illegally", assembling "illegal" genes encoding an odd number of cysteines. However, this phenomenon only seems to be a major problem for "incomplete" gene-assembling ligation reactions, in which one or both of the TGT and TGC block classes are absent. As the four block classes preferentially ligate together as designed (this had been verified by sequencing [Section 4.3.1]), the strategy for the construction of the DNA library is not flawed (Figure 3.3).

### 4.3.3 Gene assembling ligation reactions utilising low amounts of DNA

Ligations 9 and 10 [N-CAPA, TGC4, TGT4, C-CAPA] (Section 4.2.6) contained the same four species of block at the same ratio, the only difference being that ligation 10 contained 25 times less DNA. Amplification of both ligations with primers specific for the N/C-CAPA blocks gave identical ladders of bands (Figure 4.30, lanes 2 and 3, respectively) diagnostic of correct gene formation (Construct I:60bp, II:90bp, III:120bp, IV:150bp). It could be concluded therefore, that the blocks in both ligations had ligated together as designed, and that the different amounts of DNA in both ligation reactions had made no difference to the outcome of the ligation reaction or the subsequent PCR amplification. This result determined that much lower amounts of DNA could be used in subsequent ligation reactions, the significance of which is explained in Section 4.3.5.

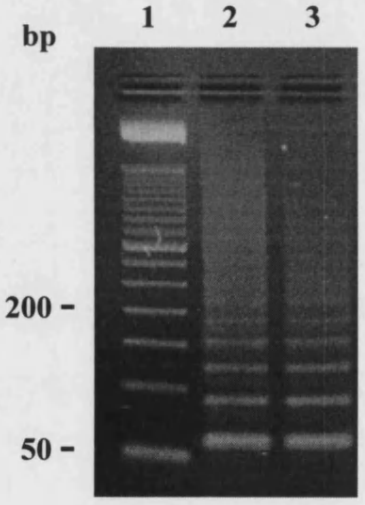



**Figure 4.30** Low amounts of DNA can be used in gene-assembling ligation reactions. Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 9 (N-CAPA, TGC4, TGT4, C-CAPA) which had a molar ratio of 25:150:100:25 pmol, respectively. Lane 3: PCR amplification of ligation 10 (N-CAPA, TGC4, TGT4, C-CAPA) which had a molar ratio of 1:6:4:1 pmol, respectively.

### 4.3.4 Incorporation of 6 bp and 9 bp dsDNA blocks

Amplification of ligation 11 (N-CAPA, TGC2, TGT4, C-CAPA) and ligation 12 (N-CAPA, TGC3, TGT4, C-CAPA) (Section 4.2.7) with primers specific for the N/C-CAPA blocks, gave ladders of bands diagnostic of correct gene formation (Figure 4.31, lanes 2 and 4).

The PCR products were subsequently cloned into pCR3. Of 20 screened transformants (10 for each ligation), seven inserts were sequenced (Figures 4.32 and 4.33). All four inserts from ligation 11 (N-CAPA, TGC2, TGT4, C-CAPA) were "legal" genes (Constructs I:1, II:1, III:1, IV:1) proving the incorporation of the 6 bp TGC2 block. Two inserts from ligation 12 (N-CAPA, TGC3, TGT4, C-CAPA) were "legal" genes (Constructs I:1, II:1) ratifying the incorporation of the 9 bp TGC3 block. However, the third insert was a "double-construct" in which two genes had ligated together because of incomplete *Xba* I digestion (Section 4.3.1). Thus 6 bp and 9 bp dsDNA propagating blocks, encoding spacers of 2 and 3 amino acids, respectively, are able to incorporate into the genes assembled by a ligation reaction.

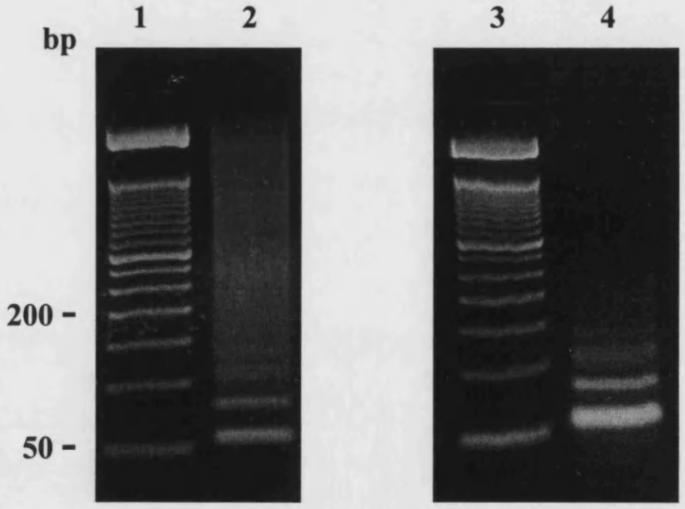



**Figure 4.31** Amplification products of assembled genes indicative of correct incorporation of the 6 bp TGC2 and 9 bp TGC3 blocks. Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 11 (N-CAPA, TGC2, TGT4, C-CAPA). Lane 3: 50 bp DNA ladder. Lane 4: PCR amplification of ligation 12 (N-CAPA, TGC3, TGT4, C-CAPA).

| Insert | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **D1** | N-CAPA | TGC2 | C-CAPA | | | | | | |
| **D2** | N-CAPA | TGC2 | TGT4 | TGC2 | C-CAPA | | | | |
| **D3** | N-CAPA | TGC2 | TGT4 | TGC2 | TGT4 | TGC2 | C-CAPA | | |
| **D4** | N-CAPA | TGC2 | TGT4 | TGC2 | TGT4 | TGC2 | TGT4 | TGC2 | C-CAPA |

**Figure 4.32** Block composition, derived from the nucleotide sequences, of four genes assembled by ligation 11 (N-CAPA, TGC2, TGT4, C-CAPA) confirming the incorporation of the 6 bp TGC2 block.

| Insert | | | | | | |
|---|---|---|---|---|---|---|
| **E1** | N-CAPA | TGC3 | C-CAPA | | | |
| **E2** | N-CAPA | TGC3 | TGT4 | TGC3 | C-CAPA | |
| **E3** | N-CAPA | TGC3 | C-CAPA | N-CAPA | TGC3 | C-CAPA |

**Figure 4.33** Block composition, derived from the nucleotide sequences, of three genes assembled by ligation 12 (N-CAPA, TGC3, TGT4, C-CAPA) confirming the incorporation of the 9 bp TGC3 block. Note that insert E3 is a "double-construct" formed due to incomplete *Xba* I digestion of the assembled genes.

### 4.3.5 Optimal means of dsDNA block formation and oligonucleotide 5′-phosphorylation

This experiment (Section 4.2.8) sought to compare the efficacy of blocks formed by lyophilisation with those formed by temperature-annealing, and to compare the efficacy of commercially 5′-phosphorylated oligonucleotides with those 5′-phosphorylated in the laboratory.

It was found that dsDNA blocks can be directly formed either by temperature-annealing (Figure 4.34, lane 8) or lyophilisation (lane 4) of complementary oligonucleotides. The TGT4 block (12 bp) appeared as a much higher band on the gel, when compared to its constituent oligonucleotides, as dsDNA molecules migrate through non-denaturing polyacrylamide gels at a slower rate than ssDNA molecules. Although both oligonucleotides were 15 bases long, there was some variation in their final migration distance as base composition and sequence can affect the migration of

ssDNA molecules through non-denaturing polyacrylamide gels (Sambrook *et al.*, 1989).

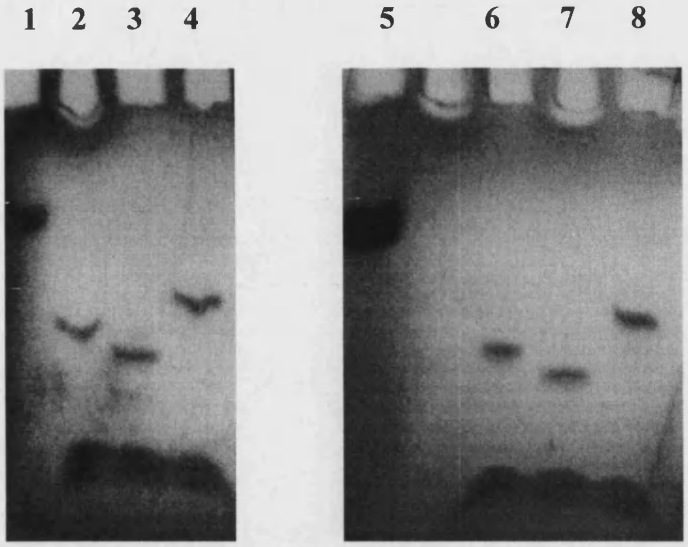



**Figure 4.34** 20% non-denaturing polyacrylamide gel displaying the TGT4 block (formed by temperature-annealing and lyophilisation) alongside its constituent oligonucleotides. Lanes 1 and 5: Xylene cyanol FF running dye. Lanes 2 and 6: 300pmol of the 15 base oligonucleotide representing the TGT4 sense strand. Lanes 3 and 7: 300pmol of the 15 base oligonucleotide representing the TGT4 anti-sense strand. Lane 4: 300pmol of the 12 bp TGT4 dsDNA block formed by lyophilisation. Lane 8: 300pmol of the 12 bp TGT4 dsDNA block formed by temperature-annealing.

Amplification of ligations 13-16 [N-CAPA, TGC4, TGT4, C-CAPA] with primers specific for the N/C-CAPA blocks, gave identical ladders of bands (Figure 4.35) diagnostic of correct gene formation (Construct I:60bp, II:90bp, III:120bp, IV:150bp). This experiment determined that blocks formed by lyophilisation were equally as good as those formed by temperature-annealing, and that oligonucleotides phosphorylated in the laboratory (Section 2.1.18) were equally as good as those which had been supplied phosphorylated. Note that the lyophilised samples of the 50pmol TGC4 and TGT4 blocks, when composed of oligonucleotides that had been phosphorylated in the laboratory, contained traces of residual glycerol.

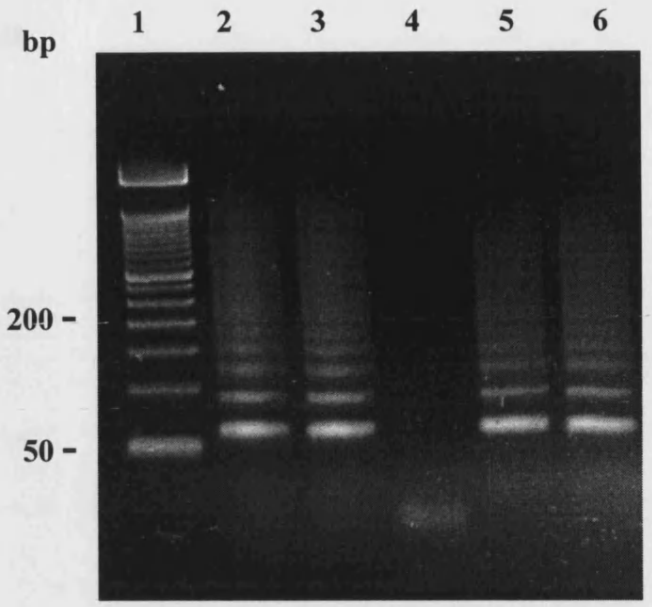



**Figure 4.35** Determination of the optimal means of dsDNA block formation and oligonucleotide 5′-phosphorylation. Ligations 13-16 [N-CAPA, TGC4, TGT4, C-CAPA] were PCR amplified, the only difference between them being in the method of dsDNA block formation (lyophilisation or temperature annealing) and the means by which the oligonucleotides comprising the propagating blocks had been 5′-phosphorylated (supplied commercially or phosphorylated in the laboratory). Lane 1: 50 bp DNA ladder. Lane 2: Blocks lyophilised, oligonucleotides supplied phosphorylated. Lane 3: Blocks temperature-annealed, oligonucleotides supplied phosphorylated. Lane 4: Negative PCR control. Lane 5: Blocks lyophilised, oligonucleotides phosphorylated in the laboratory. Lane 6: Blocks temperature-annealed, oligonucleotides phosphorylated in the laboratory.

Consequently, purely for economic reasons, all further oligonucleotides were supplied unphosphorylated (unless otherwise stated), and were phosphorylated in the laboratory (Section 2.1.18). It was also decided to form the blocks in all subsequent ligation reactions by lyophilisation (unless otherwise stated), which effectively anneals each pair of complementary oligonucleotides together at 0°C. The advantage of lyophilisation is that different sized blocks, which have significantly different annealing temperatures, can all be lyophilised simultaneously, considerably reducing the set-up time for each experiment. The only disadvantage of lyophilisation is that the maximum amount of dsDNA block that can be fully lyophilised when both complementary oligonucleotides have been phosphorylated in the laboratory (Section 2.1.18), is approximately 15pmol (Section 2.1.19).

### 4.3.6 The optimal design of the terminating blocks

Three designs of N and C terminating blocks were investigated in this chapter; N/C-CAPA (Section 4.1.2), N/C-CAPB (Section 4.1.3) and N/C-CAPC (Section 4.1.4) the different designs influencing the subsequent cloning strategy for the assembled genes. The point of this experiment (Section 4.2.9) was to determine the best design of terminating blocks to use for the construction of the DNA library.

Amplification of ligation 17 (N-CAPA, TGC4, TGT4, C-CAPA) with primers (PCR A/B) specific for the terminating blocks gave a clean ladder of bands (Figure 4.36, lane 3) diagnostic of correct gene formation (Construct I:60bp, II:90bp, III:120bp, IV:150bp). However, amplification of ligation 18 (N-CAPB, TGC4, TGT4, C-CAPB) with primers (PCRL A/B) specific for the terminating blocks gave a more complicated ladder of bands (lane 1), with the brighter ones diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp) and the fainter, intermediate bands, corresponding in size to "illegal" genes encoding an odd number of cysteines. Note that both ligations utilised the same N:TGC:TGT:C molar ratio of 2:10:10:2 pmol, respectively.

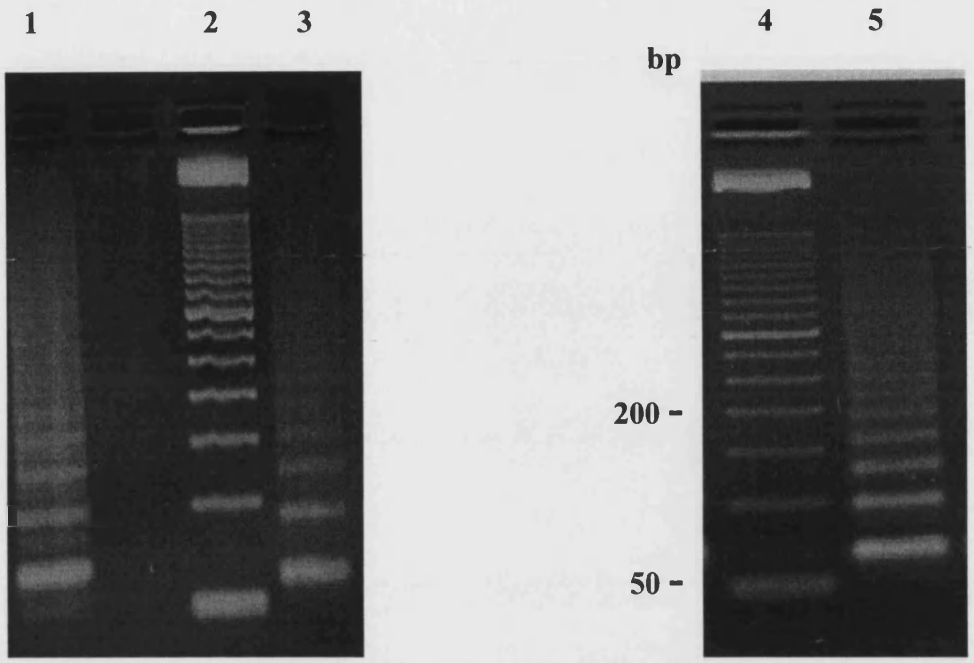



**Figure 4.36** Comparison of three different designs of terminating block. Lane 1: PCR amplification of ligation 18 (N-CAPB, TGC4, TGT4, C-CAPB) with the PCRL A/B primers. Lane 2: 50 bp DNA ladder. Lane 3: PCR amplification of ligation 17 (N-CAPA, TGC4, TGT4, C-CAPA) with the PCR A/B primers. Lane 4: 50 bp DNA ladder. Lane 5: PCR amplification of ligation 19 (N-CAPC, TGC4, TGT4, C-CAPC) with the PCRL A/B primers.

As the N, TGT, TGC and C classes of block are occasionally able to ligate together "illegally" (Section 4.3.2) every ligation reaction assembles a small proportion of "illegal" genes. As the PCRL A/B primers amplify any gene, "legal" or otherwise, containing the N/C-CAPB blocks (Section 4.1.3), both the "legal" and "illegal" genes assembled by ligation 18 (N-CAPB, TGC4, TGT4, C-CAPB) were subsequently amplified, producing the complicated ladder of bands seen. In contrast, a large proportion of the "illegal" genes assembled by ligation 17 (N-CAPA, TGC4, TGT4, C-CAPA) were not subsequently amplified, due solely to the design of the PCR A/B primers (Section 4.1.2), suggesting the superiority of the N/C-CAPA blocks.

However, the N/C-CAPA blocks were discarded in favour of the N/C-CAPB blocks, so that the first and last amino acid encoded in the spacer of each TGC block would no longer have to be fixed as valine and tyrosine, respectively. Also, negative PCR

controls (Section 4.3.1) had suggested that the PCR A/B primers were able to form primer-dimers in PCR reactions. Consequently, ligation reactions containing the N/C-CAPB blocks had to be physically optimised to avoid the assembly of "illegal" genes at the outset (Section 4.3.7).

The N/C-CAPC blocks were N/C-CAPB blocks that had been elongated by 16 bp, so that they contained *Hind* III and *Xba* I restriction sites, respectively. Therefore, the PCRL A/B primers amplify genes containing the N/C-CAPC blocks in exactly the same way as genes containing the N/C-CAPB blocks. Amplification of ligation 19 (N-CAPC, TGC4, TGT4, C-CAPC) with primers (PCRL A/B) specific for the terminating blocks gave a clean ladder of bands (lane 6) diagnostic of correct gene formation (Constructs I:66bp, II:96bp, III:126bp, IV:156bp). Note that the optimal N:TGC:TGT:C molar ratio of 1:10:6:1 pmol (Section 4.3.7) prevented the assembly of "illegal" genes.

In summary, it was decided to use the N/C-CAPC terminating blocks in the construction of the DNA library, as this gives the option of two cloning strategies for the assembled genes (Section 4.1.4), the best of which was determined in Chapter 5.

### 4.3.7 Physical optimisation of the gene-assembling ligation reactions

The objective of these experiments (Section 4.2.10) was to completely optimise the physical conditions for the gene-assembling ligation reactions.

#### Determination of ligation temperature

Amplification of ligations 20 and 21 [N-CAPB, TGC4, TGT4, C-CAPB] with primers specific for the N/C-CAPB blocks gave identical ladders of bands (Figure 4.37, lanes 2 and 3) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). The only difference between these two ligations was that the complementary oligonucleotides used to form the blocks in ligation 20 had been

denatured at 94°C for 1 minute, prior to lyophilisation. This experiment showed that this step was unnecessary, and it was subsequently omitted.

Amplification of ligation 23 (N-CAPB, TGC2, TGT6, C-CAPB) which had used a ligation temperature of 4°C, with primers specific for the N/C-CAPB blocks gave a ladder of bands (lane 5) diagnostic of correct gene formation (Construct I:60bp, II:90bp, III:120bp, IV:150bp). However, amplification of ligation 22 (which was identical to ligation 23 except that the ligation temperature was 16°C) with primers specific for the N/C-CAPB blocks gave a smear (lane 4), indicating that the 6 bp TGC2 block had not been able to incorporate properly at 16°C. This experiment determined that the 6 bp TGC2 block was not stable at 16°C, at which optimum ligation occurs using T4 DNA ligase (as recommended by the manufacturer). Consequently, all subsequent ligation reactions containing 6 bp dsDNA blocks were incubated at 4°C.

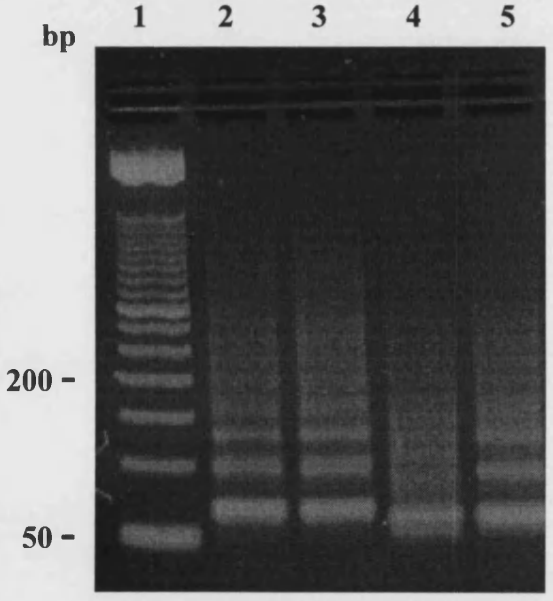



**Figure 4.37** Optimisation of ligation temperature. Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 20 (N-CAPB, TGC4, TGT4, C-CAPB) in which the complementary oligonucleotides were denatured prior to lyophilisation. Lane 3: As lane 2 except that the complementary oligonucleotides were not denatured prior to lyophilisation (ligation 21). Lane 4: PCR amplification of ligation 22 (N-CAPB, TGC2, TGT6, C-CAPB) which used a ligation temperature of 16°C. Lane 5: As lane 4 except that the ligation temperature was 4°C (ligation 23).

**Optimisation of the N:TGC:TGT:C molar ratio**

Amplification of a ligation [N-CAPB, TGC4, TGT4, C-CAPB] with primers specific for the N/C-CAPB blocks should give a ladder of bands diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). Ligations 24-37 all contained these four species of block, although the molar ratio of the blocks varied (Figure 4.38).

| Ligation | N-CAPB | TGC4 | TGT4 | C-CAPB |
|---|---|---|---|---|
| 24 | 1 | 1 | 1 | 1 |
| 25 | 1 | 2 | 2 | 1 |
| 26 | 1 | 4 | 4 | 1 |
| 27 | 1 | 4 | 2.4 | 1 |
| 28 | 1 | 2.4 | 4 | 1 |
| 29 | 1 | 4 | 4 | 1 |
| 30 | 1 | 4 | 3 | 1 |
| 31 | 1 | 4 | 2.66 | 1 |
| 32 | 1 | 4 | 2.4 | 1 |
| 33 | 1 | 4 | 2 | 1 |
| 34 | 1 | 10 | 6 | 1 |
| 35 | 2 | 10 | 6 | 2 |
| 36 | 2.5 | 10 | 6 | 2.5 |
| 37 | 4 | 10 | 6 | 4 |

**Figure 4.38** The N:TGC:TGT:C molar ratio in ligations 24-37 [N-CAPB, TGC4, TGT4, C-CAPB].

The PCR products of ligations 24-28 (Figure 4.39, lanes 2-6) clearly demonstrate that the outcome of a ligation reaction can be influenced by the molar ratio of the block classes.

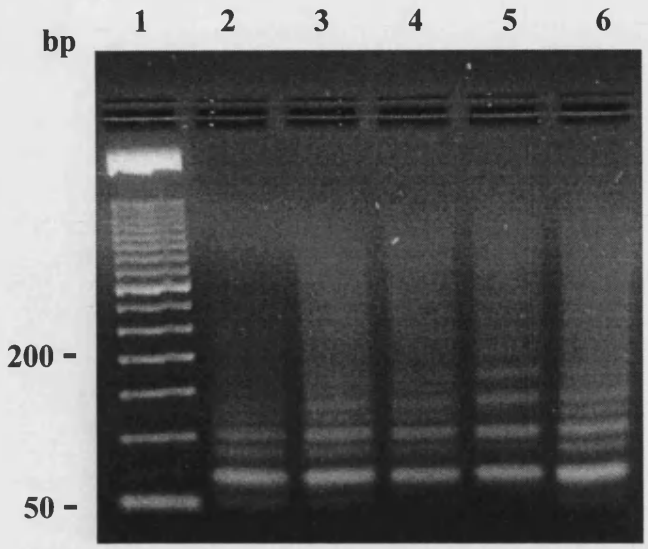



**Figure 4.39** The N:TGC:TGT:C molar ratio can influence the outcome of a gene-assembling ligation reaction. PCR amplifications of ligations 24-28 [N-CAPB, TGC4, TGT4, C-CAPB], the only difference between the ligations being in the N:TGC:TGT:C molar ratio. Lane 1: 50 bp DNA ladder. Lane 2: 1:1:1:1 molar ratio. Lane 3: 1:2:2:1 molar ratio. Lane 4: 1:4:4:1 molar ratio. Lane 5: 1:4:2.4:1 molar ratio. Lane 6: 1:2.4:4:1 molar ratio.

As the PCRL A/B primers amplify any ligation construct, "legal" or otherwise, containing the N/C-CAPB terminating blocks (Section 4.1.3) it is crucial to minimise the extent of "illegal" gene assembly during ligation reactions. The PCR products of ligations 29-33 (Figure 4.40, lanes 2-6), clearly show that the TGC:TGT molar ratio is able to influence the proportion of "legal" and "illegal" genes assembled by a ligation reaction (the presence of "illegal" genes in the ligation reaction was gauged by the extent of intermediate band formation in the PCR products, corresponding to genes encoding an odd number of cysteines [Section 4.3.6]), and that the optimum TGC:TGT molar ratio, which minimises the assembly of "illegal" genes, is 4:2.4 [10:6] (lane 5).

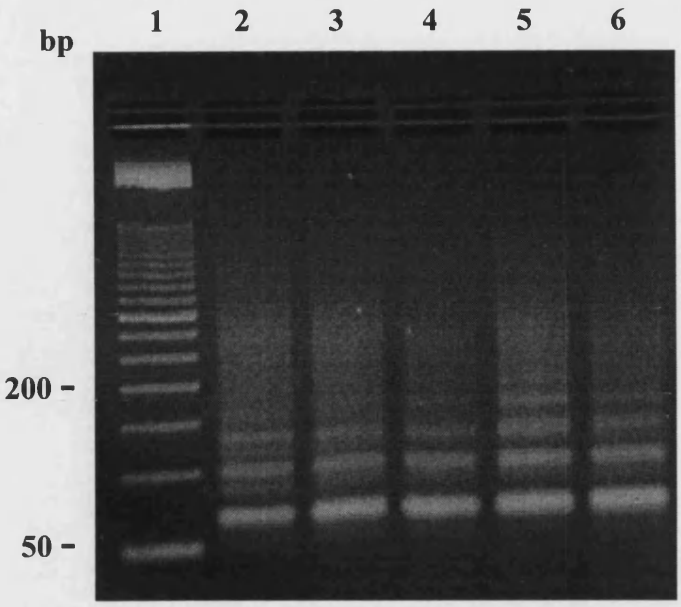



**Figure 4.40** The TGC:TGT molar ratio influences the extent of "illegal" gene assembly. PCR amplifications of ligations 29-33 [N-CAPB, TGC4, TGT4, C-CAPB], the only difference between the ligations being in the N:TGC:TGT:C molar ratio. Lane 1: 50 bp DNA ladder. Lane 2: 1:4:4:1 molar ratio. Lane 3: 1:4:3:1 molar ratio. Lane 4: 1:4:2.66:1 molar ratio. Lane 5: 1:4:2.4:1 molar ratio. Lane 6: 1:4:2:1 molar ratio.

The PCR products of ligations 34-37 (Figure 4.41, lanes 2-5) clearly show that the ratio of the terminating (N and C) to propagating (TGT and TGC) blocks is able to influence the relative proportion and average size of the products produced by a ligation reaction. Consequently, the optimum N:TGC:TGT:C molar ratio is 1:10:6:1, respectively, which produces a very even proportion of construct type with a high average size (so that constructs I-IV, encoding 2, 4, 6 and 8 cysteines are equally represented) and minimises the assembly of "illegal" genes. The efficacy of this ratio is clearly illustrated by the PCR products of ligation 34 (lane 2).

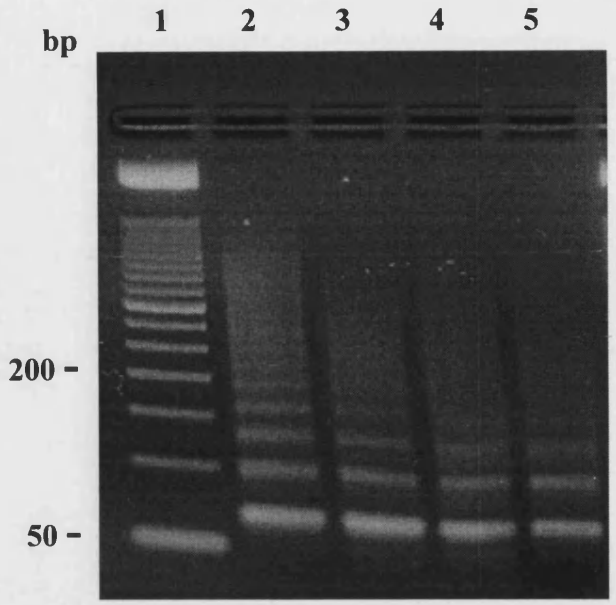



**Figure 4.41** The terminating to propagating block ratio influences the relative proportion and average size of the assembled genes. PCR amplifications of ligations 34-37 [N-CAPB, TGC4, TGT4, C-CAPB], the only difference between the ligations being in the N:TGC:TGT:C molar ratio. Lane 1: 50 bp DNA ladder. Lane 2: 1:10:6:1 molar ratio. Lane 3: 2:10:6:2 molar ratio. Lane 4: 2.5:10:6:2.5 molar ratio. Lane 5: 4:10:6:4 molar ratio.

**Efficacy of two-stage ligation reactions**

The efficacy of two-stage ligation reactions, in which the propagating blocks (TGT and TGC) are ligated together for a period of time (primary ligation) before the terminating blocks (N and C) are added (secondary ligation), was investigated in this experiment.

Amplification of ligations 38-41 [N-CAPB, TGC4, TGT4, C-CAPB] with primers specific for the N/C-CAPB blocks gave ladders of bands (Figure 4.42, lanes 2-5) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). The only difference between the four ligations was in the duration of the primary ligation (0, 30, 60 or 120 minutes) in which the propagating blocks were ligated together. This experiment determined that two-stage ligation reactions are able to increase the average size of the assembled genes and to produce an even proportion of

constructs I-IV, so that genes encoding 2, 4, 6 and 8 cysteines are equally represented. Even a primary ligation as short as 30 minutes (lane 3) produces a very even proportion of construct type (I-IV) with a high average size, when compared to an identical ligation reaction in which the propagating and terminating blocks were ligated together at the same time (lane 2). Consequently, all subsequent ligations were two-stage reactions in which the propagating blocks were ligated together for 60 minutes (primary ligation), before the terminating blocks were added (secondary ligation).

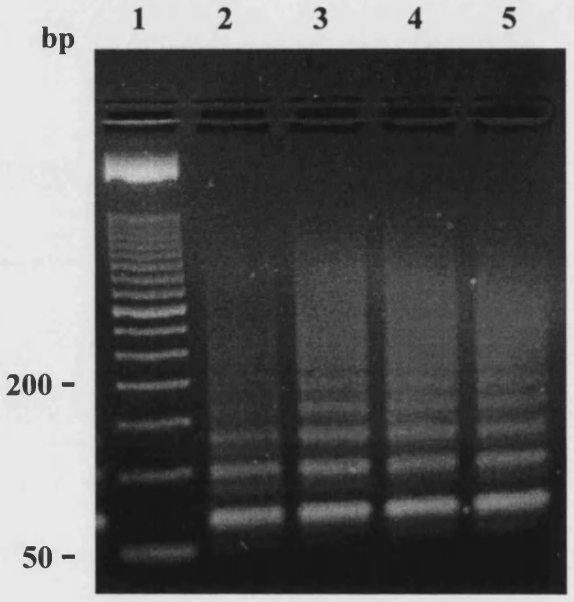



**Figure 4.42** Two stage ligation reactions influence the relative proportion and average size of the assembled genes. PCR amplifications of ligations 38-41 [N-CAPB, TGC4, TGT4, C-CAPB], the only difference between the ligations being in the duration of the primary ligation, in which the propagating blocks were ligated together. Lane 1: 50 bp DNA ladder. Lane 2: 0 minutes primary ligation. Lane 3: 30 minutes primary ligation. Lane 4: 60 minutes primary ligation. Lane 5: 120 minutes primary ligation.

### Non-preferential incorporation of different sized blocks

This experiment investigated whether T4 DNA ligase showed any size preference when incorporating different sized blocks.

Amplification of ligation 42 (N-CAPB, TGC4, TGT4, C-CAPB) with primers specific for the N/C-CAPB blocks gave a ladder of bands (Figure 4.43, lane 1) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). Similarly, amplification of ligation 44 (N-CAPB, TGC2, TGT4, C-CAPB) with primers specific for the N/C-CAPB blocks also gave a ladder of bands (lane 3) diagnostic of correct gene formation (Construct I:60bp, II:84bp, III:108bp, IV:132bp).

However, amplification of ligation 43 (N-CAPB, TGC2, TGC4, TGT4, C-CAPB), with primers specific for the N/C-CAPB blocks gave a smear (lane 2) resembling neither the PCR products of ligations 42 (lane 1) or 44 (lane 3). In ligation 43 the TGC2 and TGC4 blocks were effectively competing with one other as to which would incorporate into the ligation products. If the T4 DNA ligase had preferentially incorporated the 12 bp TGC4 block, at the expense of the 6 bp TGC2 block, the PCR products would have resembled those of ligation 42 (and vice-versa). However, the PCR products were consistent with the 6 bp TGC2 block and the 12 bp TGC4 block incorporating into the ligation products equally as well. As a consequence, it was concluded that the T4 DNA ligase showed no size preference in its incorporation of different sized blocks.

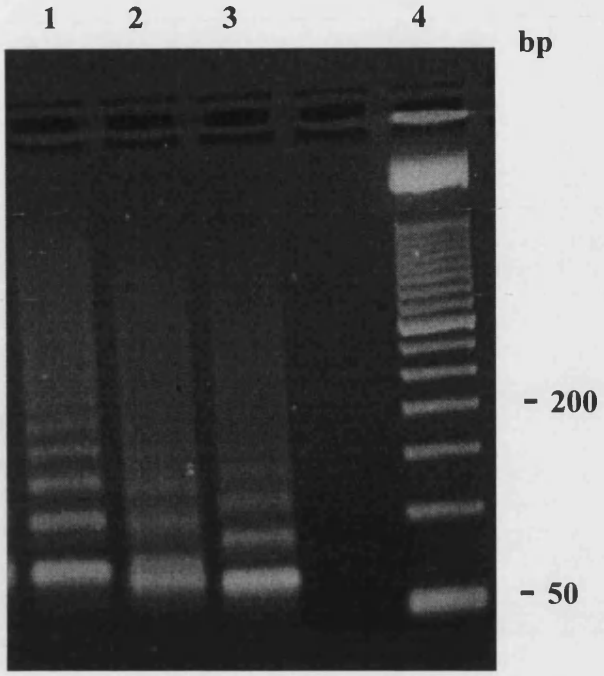



**Figure 4.43** T4 DNA ligase shows no size preference in its incorporation of different sized dsDNA blocks. Lane 1: PCR amplification of ligation 42 (N-CAPB, TGC4, TGT4, C-CAPB). Lane 2: PCR amplification of ligation 43 (N-CAPB, TGC2, TGC4, TGT4, C-CAPB). Lane 3: PCR amplification of ligation 44 (N-CAPB, TGC2, TGT4, C-CAPB). Lane 4: 50 bp DNA ladder.

### Optimum temperature and duration of ligation reactions

The physical conditions of ligation reactions had been optimised in this section, so that they were now incubated in two stages for a total of three hours at 4°C (Section 2.1.19). The aim of this experiment was to determine whether these conditions really represented the optimum temperature and duration for the ligation reactions.

The products of ligations 45-48 [N-CAPC, TGC4, TGT4, C-CAPC] appeared as identical ladders of DNA constructs (Figure 4.44, lanes 1-4). As the only difference between the four ligations was in the temperature and duration of the ligation reaction, this experiment showed that a three hour ligation at 4°C was as extensive as an overnight ligation at 16°C (at which optimal ligation occurs with the T4 DNA ligase,

as recommended by the manufacturer). Consequently, the optimal protocol for the assembly of the DNA library (Section 2.1.19) was not altered.

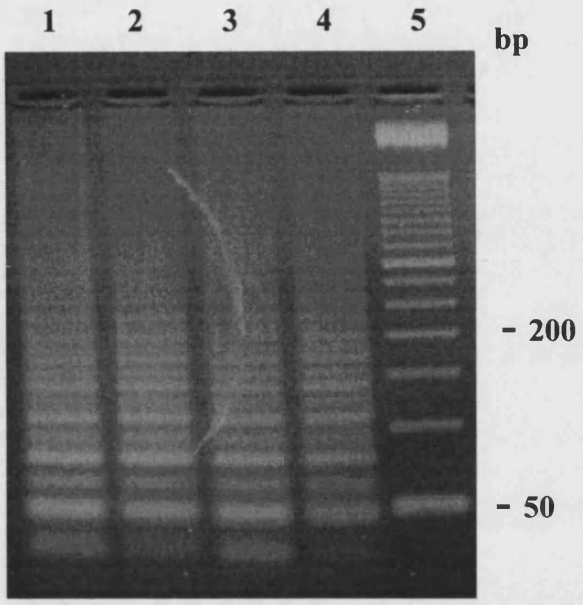



**Figure 4.44** Determination of the optimum temperature and duration of gene-assembling ligation reactions. Products of ligations 45-48 [N-CAPC, TGC4, TGT4, C-CAPC], the only difference between the ligations being in the ligation temperature and duration of the secondary ligation. Lane 1: 4°C ligation temperature, two hour secondary ligation. Lane 2: 4°C ligation temperature, overnight secondary ligation. Lane 3: 16°C ligation temperature, two hour secondary ligation. Lane 4: 16°C ligation temperature, overnight secondary ligation. Lane 5: 50 bp DNA ladder.

**An excess of terminating blocks in the secondary ligation**

Two-stage ligation reactions and the optimum N:TGC:TGT:C molar ratio of 1:10:6:1 pmol, had been combined to give the optimised protocol for the execution of subsequent ligation reactions (Section 2.1.19). This experiment sought to determine whether an excess of terminating blocks in the secondary ligation reaction, to mop up all the available propagating blocks, could improve this optimised ligation protocol even further.

Amplification of ligations 49-51 [N-CAPC, TGC4, TGT4, C-CAPC] with primers specific for the N/C-CAPC blocks, gave ladders of bands (Figure 4.45, lanes 2-4) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp).

As the only difference between the three ligation reactions was in the N:TGC:TGT:C molar ratio, this experiment determined that adding an excess of terminating blocks to the secondary ligation reaction did not improve the results of the ligation reaction. In fact, the deviation from the optimum N:TGC:TGT:C molar ratio, which had been used for ligation 49 (1:10:6:1) (lane 2), detrimentally influenced the relative proportion and average size of the products produced by ligations 50 (4:10:6:4) (lane 3) and 51 (10:10:6:10) (lane 4). Consequently, all subsequent ligation reactions used the optimum N:TGC:TGT:C molar ratio of 1:10:6:1 pmol, and no excess of terminating blocks was added to the secondary ligations (i.e. the protocol for the assembly of the DNA library [Section 2.1.19] was not altered).

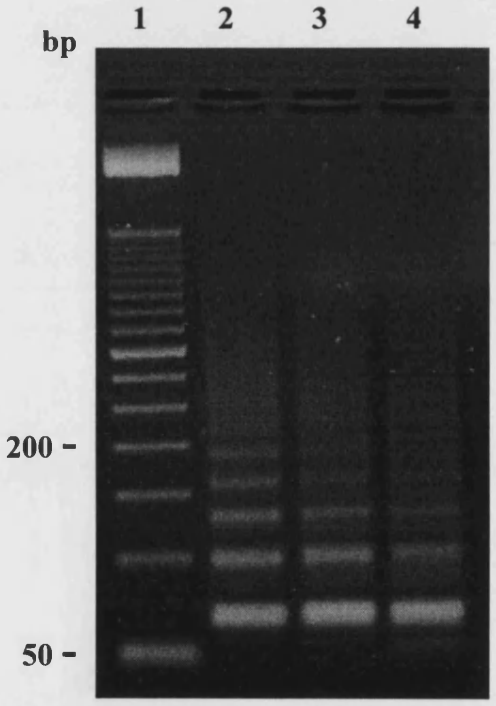



**Figure 4.45** An excess of terminating blocks in the secondary ligation reaction detrimentally influences the relative proportion and average size of the assembled genes. PCR amplifications of ligations 49-51 [N-CAPC, TGC4, TGT4, C-CAPC], the only difference between the ligations being in the N:TGC:TGT:C molar ratio. Lane 1: 50 bp DNA ladder. Lane 2: 1:10:6:1 molar ratio. Lane 3: 4:10:6:4 molar ratio. Lane 4: 10:10:6:10 molar ratio.

**Summary**

This section successfully optimised the physical conditions for the gene-assembling ligation reactions (Section 2.1.19).

### 4.3.8 Non-incorporation of 3 bp dsDNA blocks

The purpose of this experiment (Section 4.2.11) was to determine whether the 3 bp TGC1 block is able to incorporate into the genes assembled by a ligation reaction.

Amplification of ligation 52 (N-CAPB, TGC4, TGT4, C-CAPB) with primers specific for the N/C-CAPB blocks gave a ladder of bands (Figure 4.46, lane 2) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). This result served as a control by demonstrating correct gene assembly by the four classes of block at 0.5°C.

Amplification of ligations 53-55 [N-CAPB, TGC1, TGT4, C-CAPB] (which used ligation temperatures of 0.5°C, 2°C and 4°C, respectively) with primers specific for the N/C-CAPB blocks gave identical ladders of bands diagnostic of non-incorporation of the TGC1 block (lanes 3-5). In each case the bands were identical to the PCR products of ligation 56 (N-CAPB, TGT4, C-CAPB) (lane 6). Had the TGC1 block incorporated as designed, a different ladder of bands would have been produced, with the four smallest representing the first four types of construct (I:57bp, II:78bp, III:99bp, IV:120bp).

Amplification of ligation 56 (N-CAPB, TGT4, C-CAPB) which used a ligation temperature of 4°C, with primers specific for the N/C-CAPB blocks gave a ladder of bands diagnostic of "illegal" TGT4 self-polymerisation (lane 6) the bands corresponding in size to genes encoding 1, 2, ... cysteine residues. In theory, no genes should have been assembled at all, although Section 4.3.2 had determined that these classes of block can ligate together "illegally" during "incomplete" ligation reactions

such as this, and Section 4.3.6 had determined that these "illegal" genes can be PCR amplified with the PCRL A/B primers.

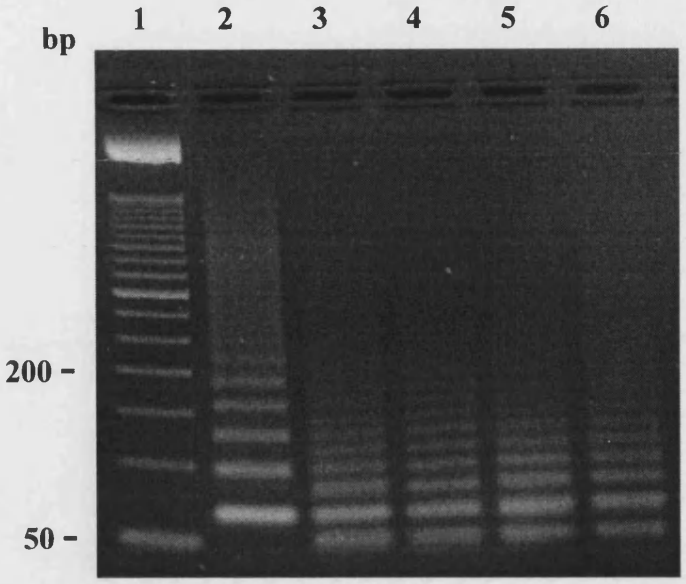



**Figure 4.46** Non-incorporation of the 3 bp TGC1 block. Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 52 (N-CAPB, TGC4, TGT4, C-CAPB) illustrating correct gene assembly at 0.5°C. Lanes 3, 4, 5: PCR amplification of ligations 53-55 [N-CAPB, TGC1, TGT4, C-CAPB] illustrating incorrect "illegal" gene assembly at 0.5°C, 2°C and 4°C. Lane 6: PCR amplification of ligation 56 (N-CAPB, TGT4, C-CAPB) illustrating "illegal" gene assembly at 4°C.

In summary, the TGC1 block had not been able to incorporate at 0.5°C, 2°C or 4°C. Consequently, the remaining N-CAPB, TGT4 and C-CAPB blocks in these reactions had ligated together "illegally" by default, as demonstrated by ligation 56 (which only contained these blocks). The 3 bp TGC1 block, was not able to incorporate, as the relative free energy of helix formation is positive (i.e. energetically impossible) for dsDNA fragments of ≤ 4 bp (Pörschke, 1977).

This experiment determined that TGT and TGC propagating blocks with spacers encoding a single amino acid are unstable. Consequently, TGC[ ] and TGT[ ] classes of block (Figure 4.1) will have to be used to cater for encoded cysteine arrangements involving two cysteines separated by a single amino acid [CXC] (as

well as arrangements involving two cysteines together [CC]) to ensure that all encoded cysteine arrangements are possible in the constructed DNA library.

### 4.3.9 Incorporation of the TGT2-6 and TGC2-6 propagating blocks

Amplification of ligations 57-66 (Section 4.2.12) with primers specific for the terminating blocks, gave ladders of bands (Figure 4.47) diagnostic of correct gene formation, each band corresponding in size to a gene encoding an even number of cysteines, with the four smallest representing constructs I-IV (Figure 4.48). Consequently, this experiment determined that the TGT2-6 and TGC2-6 propagating blocks (encoding spacers of 2-6 amino acids) are able to incorporate into the genes assembled by a ligation reaction at 4°C.

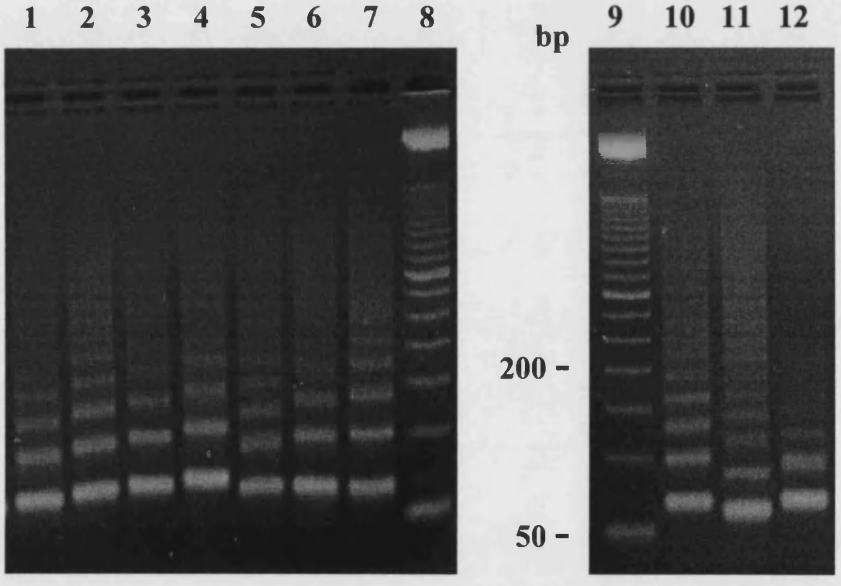



**Figure 4.47** Amplification products of ligation reactions indicative of correct incorporation of the TGT2-6 and TGC2-6 blocks. Only the propagating block compositions of the different gene-assembling ligation reactions are detailed. Lane 1: TGC3 and TGT4. Lane 2: TGC4 and TGT4. Lane 3: TGC5 and TGT4. Lane 4: TGC6 and TGT4. Lane 5: TGC4 and TGT3. Lane 6: TGC4 and TGT5. Lane 7: TGC4 and TGT6. Lane 8: 50 bp DNA ladder. Lane 9: 50 bp DNA ladder. Lane 10: TGC4 and TGT4. Lane 11: TGC2 and TGT4. Lane 12: TGC4 and TGT2.

| Ligation | Species of TGC block | Species of TGT block | Construct I (bp) | Construct II (bp) | Construct III (bp) | Construct IV (bp) |
|---|---|---|---|---|---|---|
| **57** | TGC3 | TGT4 | 63 | 90 | 117 | 144 |
| **58** | TGC4 | TGT4 | 66 | 96 | 126 | 156 |
| **59** | TGC5 | TGT4 | 69 | 102 | 135 | 168 |
| **60** | TGC6 | TGT4 | 72 | 108 | 144 | 180 |
| **61** | TGC4 | TGT3 | 66 | 93 | 120 | 147 |
| **62** | TGC4 | TGT5 | 66 | 99 | 132 | 165 |
| **63** | TGC4 | TGT6 | 66 | 102 | 138 | 174 |
| **64** | TGC4 | TGT4 | 66 | 96 | 126 | 156 |
| **65** | TGC2 | TGT4 | 60 | 84 | 108 | 132 |
| **66** | TGC4 | TGT2 | 66 | 90 | 114 | 138 |

**Table 4.48** Composition of ligations 57-66 by species of propagating block, and the sizes of the first four types of construct (I-IV) assembled by these ligations indicating correct incorporation of the TGT2-6 and TGC2-6 blocks.

### 4.3.10 The TGC[ ] and TGT[ ] classes of block

This experiment (Section 4.2.13) investigated whether the TGC[ ] and TGT[ ] classes of block were able to incorporate themselves correctly into the genes assembled by a ligation reaction.

Amplification of ligation 67 (N-CAPC, TGC[CC]5, TGC3, C-CAPC) and ligation 68 (N-CAPC, TGT[CC]5, TGC3, C-CAPC) with primers specific for the N/C-CAPC blocks gave identical ladders of bands (Figure 4.49, lanes 2 and 3) diagnostic of correct gene formation (Construct I:63bp, II:84bp, III:105bp, IV:126bp). These results suggested that the TGC[ ] and TGT[ ] classes of block (represented by TGC[CC]5 and TGT[CC]5, respectively) are able to incorporate into the genes assembled by a ligation reaction as intended. Note that the TGC[ ] and TGT[ ] block classes are able to self-polymerise by design.

Amplification of ligation 69 (N-CAPC, TGC4, TGT4, C-CAPC) with primers specific for the N/C-CAPC blocks gave a ladder of bands (lane 5) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). This result served as a

control by demonstrating correct gene assembly by the N, TGT, TGC and C block classes.

Amplification of ligation 70 (N-CAPC, TGC4, TGT4, TGC[CC]5, C-CAPC), ligation 71 (N-CAPC, TGC4, TGT4, TGT[CC]5, C-CAPC) and ligation 72 (N-CAPC, TGC4, TGT4, TGC[CC]5, TGT[CC]5, C-CAPC) with primers specific for the N/C-CAPC blocks gave very complicated ladders of bands (lanes 6, 7, 8, respectively) suggesting correct gene formation, as the propagating blocks in these reactions could ligate together in many different combinations.

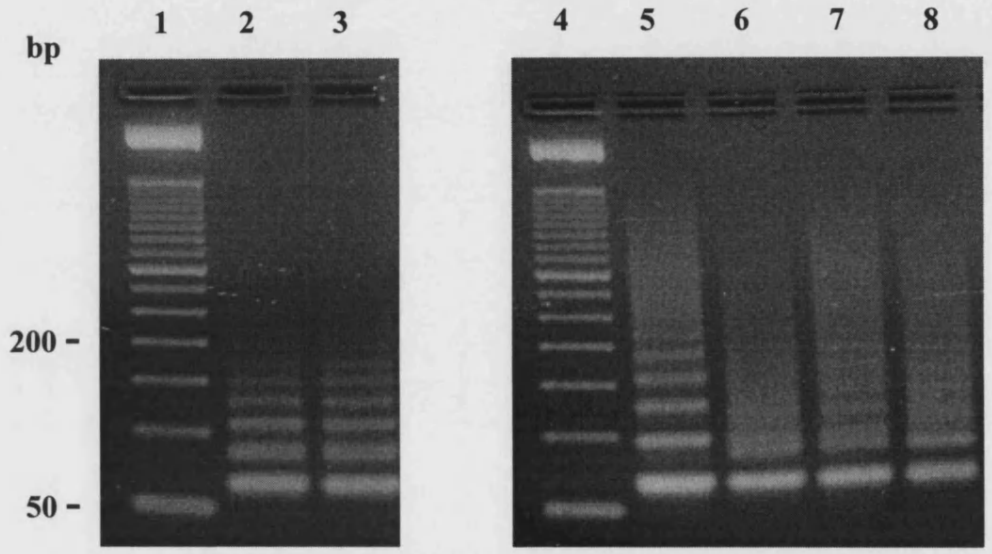



**Figure 4.49** Amplification products of ligation reactions suggestive of correct incorporation of the TGC[ ] and TGT[ ] classes of propagating block. Only the propagating block compositions of the different gene-assembling ligation reactions are detailed. Lane 1: 50 bp DNA ladder. Lane 2: TGC[CC]5 and TGC3. Lane 3: TGT[CC]5 and TGC3. Lane 4: 50 bp DNA ladder. Lane 5: TGC4 and TGT4. Lane 6: TGC4, TGT4 and TGC[CC]5. Lane 7: TGC4, TGT4 and TGT[CC]5. Lane 8: TGC4, TGT4, TGC[CC]5 and TGT[CC]5.

In order to unequivocally prove the correct incorporation of the TGC[ ] and TGT[ ] block classes, the genes assembled by ligation 72 (N-CAPC, TGC4, TGT4, TGC[CC]5, TGT[CC]5, C-CAPC) were cloned into pCR3 and 5 inserts were sequenced (Figure 4.50). Four inserts were "legal" genes encoding an even number of cysteines (Constructs II:3, III:1). However, one insert was a "double-construct", in

which two genes had ligated together because a *Xba* I digested C-CAPC block and *Hin*d III digested N-CAPC block had both become blunt-ended and had subsequently ligated together. This experiment determined that the N, TGT, TGC, TGC[ ], TGT[ ] and C classes of block are able to ligate together as designed.

| Insert | | | | | | | |
|---|---|---|---|---|---|---|---|
| **F1** | N-CAPC | TGC[CC]5 | TGC4 | C-CAPC | | | |
| **F2** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC |
| **F3** | N-CAPC | TGC4 | TGT[CC]5 | C-CAPC | | | |
| **F4** | N-CAPC | TGC[CC]5 | TGC4 | C-CAPC | | | |
| **F5** | N-CAPC | TGC4 | C-CAPC | N-CAPC | TGC[CC]5 | TGC4 | C-CAPC |

**Figure 4.50** Block composition, derived from the nucleotide sequences, of five genes assembled by ligation 72 (N-CAPC, TGC4, TGT4, TGC[CC]5, TGT[CC]5, C-CAPC) demonstrating correct incorporation of the TGC[ ] and TGT[ ] classes of supplementary block. Note that insert F5 is a "double-construct".

### 4.3.11 Elucidation of the "double-band" phenomenon

Amplification of a ligation reaction containing a single species of each of the N, TGT, TGC and C classes of block, using primers specific for the N and C terminating blocks, would be expected to produce a ladder of discretely sized DNA constructs, the four smallest representing constructs I-IV. Occasionally, if the PCR products of such a ligation reaction were separated electrophoretically for too long, each discrete band of DNA seemed to separate into two. This experiment (Section 4.2.14) aimed to elucidate the nature of this "double-band" phenomenon.

Amplification of ligation 73 (N-CAPB, TGC4, TGT4, C-CAPB) with primers specific for the N/C-CAPB blocks gave a ladder of bands (Figure 4.51, lane 2) diagnostic of correct gene formation (Construct I:66bp, II:96bp, III:126bp, IV:156bp). However, after further electrophoresis, the bands seem to separate into two, forming "double-bands" (lane 4).

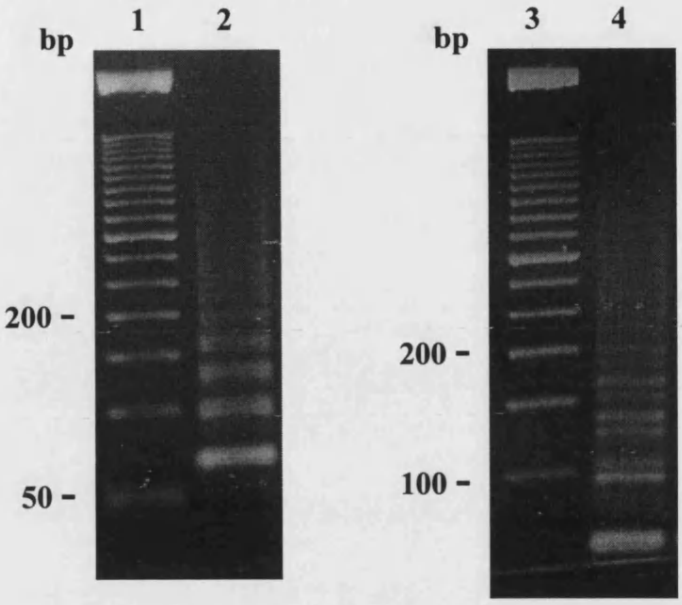



**Figure 4.51** Illustration of the "double-band" phenomenon. Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 73 (N-CAPB, TGC4, TGT4, C-CAPB) demonstrating correct gene assembly. Lanes 3 and 4: Further electrophoresis of lanes 1 and 2, displaying the "double-band" phenomenon.

As expected, amplification of the four cloned inserts representing constructs I-IV of ligation 1 (N-CAPA, TGC4, TGT4, C-CAPA) with primers specific for the N/C-CAPA blocks, gave discrete bands of 60 bp (construct I), 90 bp (construct II), 120 bp (construct III) and 150 bp (construct IV)(Figure 4.52, lanes 2-5). When combinations of these PCR products were mixed together and untreated (i.e. not denatured and re-annealed) the bands appeared as normal (lanes 6,8,10,12). However, when the same combinations of PCR products were denatured and allowed to re-anneal together, the "double-bands" appeared (lanes 7,9,11,13). This experiment determined that the "double-bands" were probably caused by the differently sized PCR products misannealing together. As there is so much complementarity between the differently sized PCR products, if they are denatured and allowed to re-anneal together, they can misanneal, the misannealed constructs appearing as the "double-bands". Thus the "double-bands" ought to pose no problem to the construction of the DNA library: they are not "illegal" constructs, only artefacts probably resulting from the high degree of complementarity found in the PCR products of such simple ligation reactions.

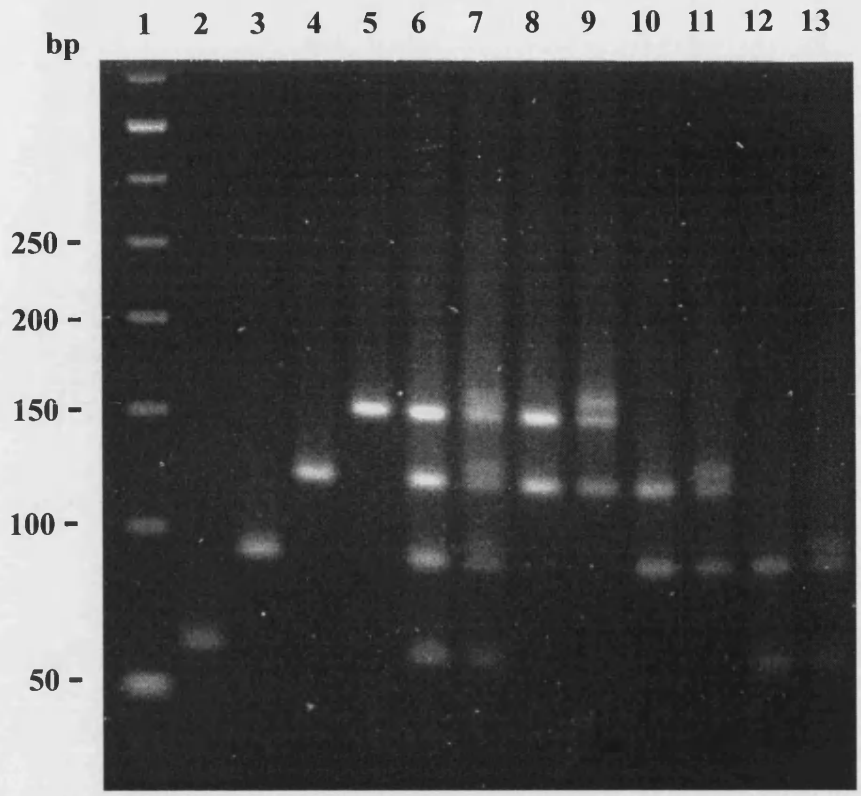



**Figure 4.52** Elucidation of the "double-band" phenomenon. Constructs I-IV assembled by ligation (N-CAPA, TGC4, TGT4, C-CAPA) were independently PCR amplified, and combinations of the PCR products were mixed together in duplicate. One reaction was denatured/re-annealed whilst the other was left untreated at room temperature. Lane 1: 50 bp DNA ladder. Lane 2: Construct I. Lane 3: Construct II. Lane 4: Construct III. Lane 5: Construct IV. Lane 6: Constructs I, II, III and IV (untreated). Lane 7: Constructs I, II, III and IV (denatured/re-annealed). Lane 8: Constructs III and IV (untreated). Lane 9: Constructs III and IV (denatured/re-annealed). Lane 10: Constructs II and III (untreated). Lane 11: Constructs II and III (denatured/re-annealed). Lane 12: Constructs I and II (untreated). Lane 13: Constructs I and II (denatured/re-annealed).

## 4.4 SUMMARY

The work described in this chapter successfully optimised every aspect of the process by which the N, TGT, TGC, TGC[ ], TGT[ ] and C classes of dsDNA block are to construct the DNA library.

Annealing pairs of complementary oligonucleotides by lyophilisation was determined to be the optimal means of procuring the dsDNA blocks, as precise concentrations of

extremely pure blocks can be formed, and since no purification steps are necessary, the blocks can be used immediately in subsequent gene-assembling ligation reactions. Also, since lyophilisation effectively anneals the complementary oligonucleotides at 0°C, different sized blocks with significantly different annealing temperatures can all be lyophilised simultaneously, considerably reducing the set-up time for each experiment. It was also decided, purely for economic reasons, to 5′-phosphorylate the oligonucleotides comprising the blocks in the laboratory (Section 2.1.18).

The optimum molar ratio of the N, TGC, TGT and C block classes in a ligation reaction, which can be extrapolated to include the TGC[ ] and TGT[ ] block classes, was empirically determined to be 1:10:6:1, respectively. This ratio minimises "illegal" gene assembly (see below) and ensures the production of an even proportion of constructs I-IV (so that genes encoding 2, 4, 6 and 8 cysteines are equally represented in the DNA library) by assembling genes with a high average size. The physical conditions for the gene-assembling ligation reactions were also completely optimised (Section 2.1.19).

The strategy for constructing the DNA library (Figure 3.3) is not perfect, as the three nucleotide 5′ overhangs utilised by the six block classes, are able to disobey standard base-pairing under certain conditions, leading to the assembly of "illegal" genes. However, "illegal" gene assembly seems to occur at a very low level in ligation reactions utilising the optimum N:TGC:TGT:C molar ratio.

Crucially, the smallest TGT and TGC propagating blocks that can successfully incorporate into the genes assembled by a ligation reaction are 6 bp in size (which contain spacers encoding 2 amino acids). The TGT and TGC blocks with spacers encoding 0 and 1 amino acids are non-existent and unstable, respectively. Consequently, the TGC[ ] and TGT[ ] classes of propagating block must be used in the construction of the DNA library to cater for cysteine arrangements involving two adjacent cysteines [CC] and two cysteines separated by a single amino acid [CXC].

Finally, the ability of the TGC2-6, TGT2-6, TGC[CC]5 and TGT[CC]5 species of propagating block (Section 4.1) to incorporate into the genes assembled by a ligation

reaction was verified, and the design of the N and C terminating blocks was optimised to give the N/C-CAPC terminating blocks (Section 4.1.4). These terminating blocks provide a choice of two cloning strategies, so that assembled genes can either be cloned directly into pCR3, or PCR amplified with primers which impose no sequence constraints upon the propagating blocks, and subsequently cloned into pCR3.

# CHAPTER 5

# EVALUATION OF THE STRATEGY FOR CONSTRUCTING THE DNA LIBRARY

## 5.0 INTRODUCTION

The previous two chapters had ascertained the procedure for constructing the DNA library by determining the design of the N, TGT, TGC, TGT[ ], TGC[ ] and C classes of dsDNA block, and by completely optimising the process by which these blocks are to assemble the genes constituting the DNA library.

This chapter sought to determine the optimum cloning strategy for the assembled genes, and to actually test the whole procedure by constructing a DNA library encoding a population of constrained peptides, using fourteen species of block (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC) representing all six classes (Figure 5.1). Note that such a DNA library, if successfully constructed, would encode an "incomplete" repertoire of constrained peptides, in the sense that not all cysteine arrangements would be possible (Chapter 6). The encoded peptides should also strictly be referred to as "potentially" constrained peptides, since not every cysteine arrangement will produce a constrained peptide.

| Construct | Assembled Genes (bp) | *Hind* III / *Xba* I Digestion (bp) | PCR Screen (bp) |
|---|---|---|---|
| I | 74-86 | 44-56 | 109-121 |
| II | 92-128 | 62-98 | 127-163 |
| III | 110-170 | 80-140 | 145-205 |
| IV | 128-212 | 98-182 | 163-247 |
| V | 146-254 | 116-224 | 181-289 |
| VI | 164-296 | 134-266 | 199-331 |
| VII | 182-338 | 152-308 | 217-373 |
| VIII | 200-380 | 170-350 | 235-415 |

**Figure 5.1** The theoretical minimum and maximum sizes of the genes assembled by ligation [N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC] representing constructs I-VIII. Also detailed are the sizes of the assembled genes after digestion with *Hind* III / *Xba* I, and once cloned into pCR3, after PCR screening of the recombinant DNA constructs with the Sp6 and T7 primers. Note that if the TGC[CC]5 and TGT[CC]5 supplementary blocks are omitted from the ligation the statistics are not altered.

The N/C-CAPC terminating blocks provide a choice of two cloning strategies (Section 4.1.4). Assembled genes containing these terminating blocks can either be cloned directly into pCR3 after digestion with *Hind* III / *Xba* I, or PCR amplified with the PCRL A and PCRL B primers and cloned into pCR3 after digestion of the PCR products with *Hind* III / *Xba* I. The only assembled genes that we actually wish to clone and express as the peptide repertoire, are those which encode peptides of 30 or fewer amino acids and which encode either 2, 4, 6 or 8 cysteines (i.e. constructs I-IV). Therefore, the cloning strategy must enable the assembled genes of interest to be selectively isolated, and must produce an even distribution of constructs I-IV in the recombinant DNA constructs (so that genes encoding 2, 4, 6 and 8 cysteines are equally represented in the DNA library). Note that the length of a peptide encoded by an assembled gene is defined as the number of amino acids between the outermost cysteines inclusively.

## 5.1 MATERIALS AND METHODS

### 5.1.1 Determination of the optimum cloning strategy for simple ligations

Ligation 1 (Figure 5.2) in a final volume of 100µl, was set up (Section 2.1.19). Oligonucleotides that had been supplied phosphorylated were used to form the TGC4 and TGT4 blocks so that large amounts could be formed by lyophilisation (Section 2.1.19).

| Ligation 1 | N-CAPC | TGC4 | TGT4 | C-CAPC |
|---|---|---|---|---|
| pmol | 5.5 | 55 | 33 | 5.5 |

**Figure 5.2** Composition of ligation 1 by block species.

A 0.5µl volume of ligation 1 was amplified in a *Taq* PCR reaction using the PCRL A and PCRL B primers and the programme: [94°C, 1 min; 65°C, 1 min; 72°C, 1 min] x 25; 72°C, 5 min. The PCR products were alcohol precipitated (Section 2.1.8), digested with *Hin*d III / *Xba* I (Section 2.1.11), separated from the released ends on a 1.5% LMP agarose gel (Section 2.1.7) and resuspended in 20µl water. A 5.5µl volume of the *Hin*d III / *Xba* I digested PCR products was then ligated into pCR3, transformed into *E. coli* and PCR screened for the presence of insert as detailed below.

The products of ligation 1 (in a volume of 99µl) were alcohol precipitated (Section 2.1.8), digested with *Hin*d III / *Xba* I (Section 2.1.11), and separated on a 3% TBE agarose gel (Section 2.1.6) at 85V for 45 minutes. A piece of 3% TBE agarose gel containing all the DNA constructs sized between 50-150 bp was subsequently excised using a clean razor blade. The excised gel piece was rotated 180° and inserted into a pre-set 1.5% LMP agarose gel (Section 2.1.7) from which an identically sized piece of gel had been removed. It was necessary to use both buffer systems (TAE and TBE) in this experiment as 3% TAE agarose gels quench the fluorescence of ethidium bromide, making visualisation of the ligation products difficult, and β-agarase I purification from LMP agarose gels demands the use of TAE buffer. A 100µl volume

of molten 1.5% LMP agarose gel, set after a 15 minute incubation at 4°C, firmly fixed the inserted piece of 3% TBE agarose gel in place. The 1.5% LMP agarose gel was then run at 40V for 45 minutes at 4°C. Rotating the inserted piece of 3% TBE agarose gel by 180° ensured that the differently sized DNA fragments migrated into the 1.5% LMP agarose gel to form a relatively compact band of DNA, which was subsequently excised using a clean razor blade (which cut approximately 2mm around the DNA smear to ensure the removal of all the DNA). The DNA fragments were extracted using β-agarase I (Section 2.1.7) and resuspended in 20μl water. A 5.5μl volume of the *Hind* III / *Xba* I digested ligation products (sized between 50-150 bp) was then ligated into 240ng (75 fmol) pCR3 (Sections 2.1.3 and 2.1.12). The recombinant DNA constructs were transformed into *E. coli* TOP 10F` cells (Section 2.1.13) and PCR screened for the presence of insert in a *Taq* PCR reaction using the Sp6 and T7 primers and the programme: 94°C, 5 min; [94°C, 1 min; 55°C, 1 min; 72°C, 1 min] x 30; 72°C, 5 min. Twelve inserts were sequenced using the pCR3seq primer.

### 5.1.2 Determination of the optimum cloning strategy for multi-block ligations

Ligation 2 (Figure 5.3) in a final volume of 300μl, was set up (Section 2.1.19) as a multi-block ligation reaction, containing twelve species of block, representing the N, TGC, TGT and C classes. The optimum molar ratio of these four block classes was preserved (Figure 5.4). Ligation 2 was cloned with both cloning strategies (Section 5.1.1) except that for the second strategy a 275μl volume of ligation 2 was alcohol precipitated, and after digestion with *Hind* III / *Xba* I, the DNA fragments sized between 44-150 bp were isolated and cloned into pCR3. A single insert was sequenced using the pCR3seq primer.

| Ligation 2 | pmol |
|---|---|
| N-CAPC | 6 |
| TGC2 | 12 |
| TGC3 | 12 |
| TGC4 | 12 |
| TGC5 | 12 |
| TGC6 | 12 |
| TGT2 | 7.2 |
| TGT3 | 7.2 |
| TGT4 | 7.2 |
| TGT5 | 7.2 |
| TGT6 | 7.2 |
| C-CAPC | 6 |

**Figure 5.3** Composition of ligation 2 by block species.

| Ligation 2 | pmol | Ratio |
|---|---|---|
| N | 6 | 1 |
| TGC | 60 | 10 |
| TGT | 36 | 6 |
| C | 6 | 1 |

**Figure 5.4** Composition of ligation 2 by block class.

### 5.1.3 Preliminary construction of a DNA library encoding a population of constrained peptides

Ligation 3 (Figure 5.5) in a final volume of 300µl, was set up (Section 2.1.19) as a multi-block ligation reaction, containing twelve species of block, representing the N, TGT, TGC and C classes. The optimum ratio of these four classes of block was preserved (Figure 5.6). The products of ligation 3 were cloned directly into pCR3 after digestion with *Hin*d III / *Xba* I (Section 5.1.1) except that a 300µl volume of ligation 3 was alcohol precipitated, and after digestion with *Hin*d III / *Xba* I, the DNA fragments sized between 62-122 bp were isolated and cloned into pCR3. Five inserts were sequenced using the pCR3seq primer.

| Ligation 3 | pmol |
|---|---|
| N-CAPC | 6 |
| TGC2 | 12 |
| TGC3 | 12 |
| TGC4 | 12 |
| TGC5 | 12 |
| TGC6 | 12 |
| TGT2 | 7.2 |
| TGT3 | 7.2 |
| TGT4 | 7.2 |
| TGT5 | 7.2 |
| TGT6 | 7.2 |
| C-CAPC | 6 |

**Figure 5.5** Composition of ligation 3 by block species.

| Ligation 3 | pmol | Ratio |
|---|---|---|
| N | 6 | 1 |
| TGC | 60 | 10 |
| TGT | 36 | 6 |
| C | 6 | 1 |

**Figure 5.6** Composition of ligation 3 by block class.

### 5.1.4 Construction of a DNA library encoding a population of constrained peptides: Part I

Ligation 4 (Figure 5.7) in a final volume of 350µl, was set up (Section 2.1.19) as a multi-block ligation reaction, containing fourteen species of block, representing the N, TGT, TGC, TGT[ ], TGC[ ] and C classes (Figure 5.8). The amount of the TGT[ ] and TGC[ ] classes of block in the ligation reaction was chosen so that each would incorporate into the ligation products, theoretically, as often as a species of TGT block. It is possible to represent the composition of ligation 4 in terms of only the N, TGT, TGC and C classes of block (Section 4.2.13). The optimum ratio of these four classes of block was preserved in ligation 4 (Figure 5.9).

The products of ligation 4 were cloned directly into pCR3 after digestion with *Hin*d III / *Xba* I (Section 5.1.1) except that a 350µl volume of ligation 3 was alcohol precipitated, and after digestion with *Hin*d III / *Xba* I the DNA fragments sized between 62-122 bp were isolated and cloned into pCR3. Twenty inserts were sequenced using the pCR3seq primer.

| Ligation 4 | pmol |
|---|---|
| N-CAPC | 6 |
| TGC2 | 9.9 |
| TGC3 | 9.9 |
| TGC4 | 9.9 |
| TGC5 | 9.9 |
| TGC6 | 9.9 |
| TGT2 | 5.2 |
| TGT3 | 5.2 |
| TGT4 | 5.2 |
| TGT5 | 5.2 |
| TGT6 | 5.2 |
| TGC[CC]5 | 5.2 |
| TGT[CC]5 | 5.2 |
| C-CAPC | 6 |

**Figure 5.7** Composition of ligation 4 by block species.

| Ligation 4 | pmol |
|---|---|
| N | 6 |
| TGC | 49.5 |
| TGT | 26 |
| TGC[ ] | 5.2 |
| TGT[ ] | 5.2 |
| C | 6 |

**Figure 5.8** Composition of ligation 4 by block class.

| Ligation 4 | pmol | Ratio |
|---|---|---|
| N | 6 | 1 |
| TGC | 60 | 10 |
| TGT | 36 | 6 |
| C | 6 | 1 |

**Figure 5.9** Refined composition of ligation 4 by block class.

### 5.1.5 Construction of a DNA library encoding a population of constrained peptides: Part II

Ligation 5 (Figure 5.10) in a final volume of 350µl, was set up (Section 2.1.19) as a multi-block ligation reaction, containing fourteen species of block, representing the N, TGT, TGC, TGT[ ], TGC[ ] and C classes (Figure 5.11). The amounts of the TGT and TGC block species were staggered in order to increase the proportion of smaller blocks in the DNA library (the staggering is represented as a percentage of the median value). The amount of the TGT[ ] and TGC[ ] classes of block in the ligation reaction was chosen so that each would incorporate into the ligation products, theoretically, as often as the TGT4 block. It is possible to represent the composition of ligation 5 in terms of only the N, TGT, TGC and C block classes (Section 4.2.13). The optimum ratio of these four block classes was preserved in ligation 4 (Figure 5.12).

The concentrations of all the oligonucleotides used in ligation 5 were verified on a spectrophotometer (Section 2.1.5) and fresh stocks of phosphorylated oligonucleotides were prepared (Section 2.1.18).

The products of ligation 5 were cloned directly into pCR3 after digestion with *Hin*d III / *Xba* I (Section 5.1.1) except that a 350µl volume of ligation 3 was alcohol precipitated, and after digestion with *Hin*d III / *Xba* I, the DNA fragments sized between 62-150 bp were isolated and subsequently cloned into pCR3. Once these recombinant DNA constructs had been transformed into *E. coli*, the colonies that were subsequently PCR screened were all a minimum of 1cm away from each other on the transformation plate. Eighteen inserts were sequenced using the pCR3seq primer.

| Ligation 5 | pmol | Staggering |
|---|---|---|
| N-CAPC | 6 | |
| TGC2 | 12.4 | 125% |
| TGC3 | 11.2 | 112.5% |
| TGC4 | 9.9 | 100% |
| TGC5 | 8.7 | 87.5% |
| TGC6 | 7.5 | 75% |
| TGT2 | 6.4 | 125% |
| TGT3 | 5.8 | 112.5% |
| TGT4 | 5.2 | 100% |
| TGT5 | 4.5 | 87.5% |
| TGT6 | 3.9 | 75% |
| TGC[CC]5 | 5.2 | |
| TGT[CC]5 | 5.2 | |
| C-CAPC | 6 | |

**Figure 5.10** Composition of ligation 5 by block species. The staggering of the TGT and TGC blocks is represented as a percentage of the median value.

| Ligation 5 | pmol |
|---|---|
| N | 6 |
| TGC | 49.7 |
| TGT | 25.8 |
| TGC[ ] | 5.2 |
| TGT[ ] | 5.2 |
| C | 6 |

**Figure 5.11** Composition of ligation 5 by block class.

| Ligation 5 | pmol | Ratio |
|---|---|---|
| N | 6 | 1 |
| TGC | 60 | 10 |
| TGT | 36 | 6 |
| C | 6 | 1 |

**Figure 5.12** Refined composition of ligation 5 by block class.

## 5.2 RESULTS AND DISCUSSION

### 5.2.1 Determination of the optimum cloning strategy for simple ligations

This experiment (Section 5.1.1) set out to determine the optimum cloning strategy for genes assembled by a ligation reaction. The different types of construct assembled by ligation 1 (N-CAPC, TGC4, TGT4, C-CAPC) can be easily distinguished from one another by virtue of their size (I:80bp, II:110bp, III:140bp, IV:170bp). Consequently, the proportion of construct type in the recombinant DNA constructs can be easily determined by PCR screening (I:115bp, II:145bp, III:175bp, IV:205bp).

Amplification of ligation 1 with primers specific for the terminating blocks gave a ladder of bands (Figure 5.13, lane 1) which were cloned into pCR3. Eighteen transformants were PCR screened (Figure 5.14), 16 of which gave bands corresponding to constructs I-IV (I:7, II:3, III:4, IV:2), with a bias towards construct I. One band, of approximately 265 bp, probably represented a type VI construct, whilst another band, of approximately 130 bp, probably represented an "illegal" gene encoding three cysteines.

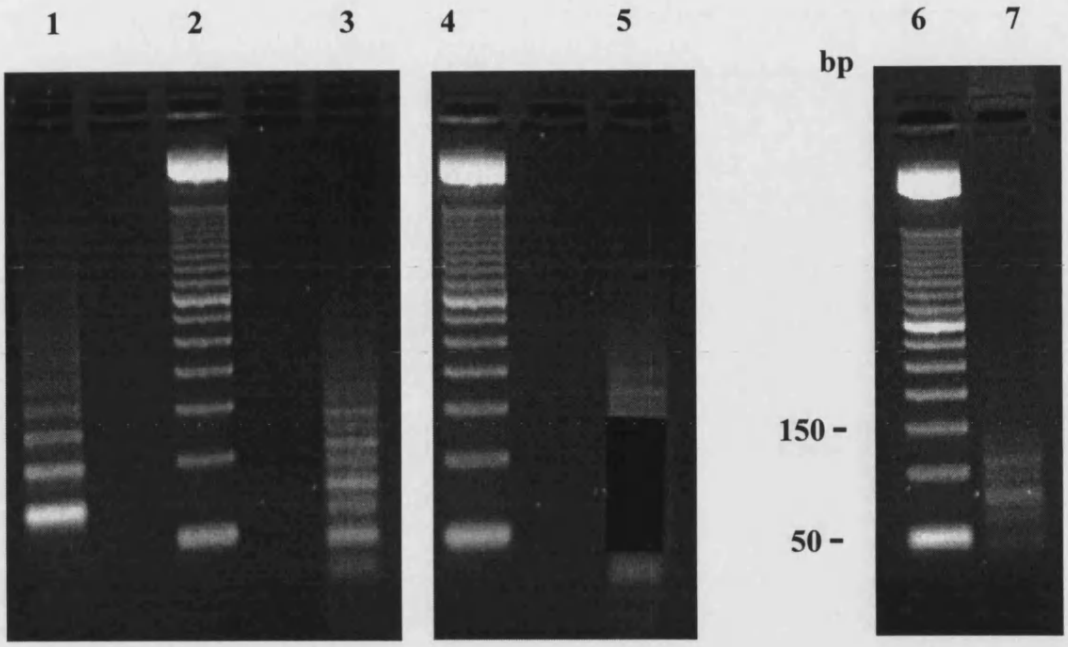



**Figure 5.13** The cloning of ligation 1 (N-CAPC, TGC4, TGT4, C-CAPC). Lane 1: PCR amplification of ligation 1. Lane 2: 50 bp DNA ladder. Lane 3: *Hin*d III / *Xba* I digested products of ligation 1. Lane 4: 50 bp DNA ladder. Lane 5: Excision of 50-150 bp DNA fragments from lane 3. Lane 6: 50 bp DNA ladder. Lane 7: Isolation of 50-150 bp DNA fragments from lane 3.

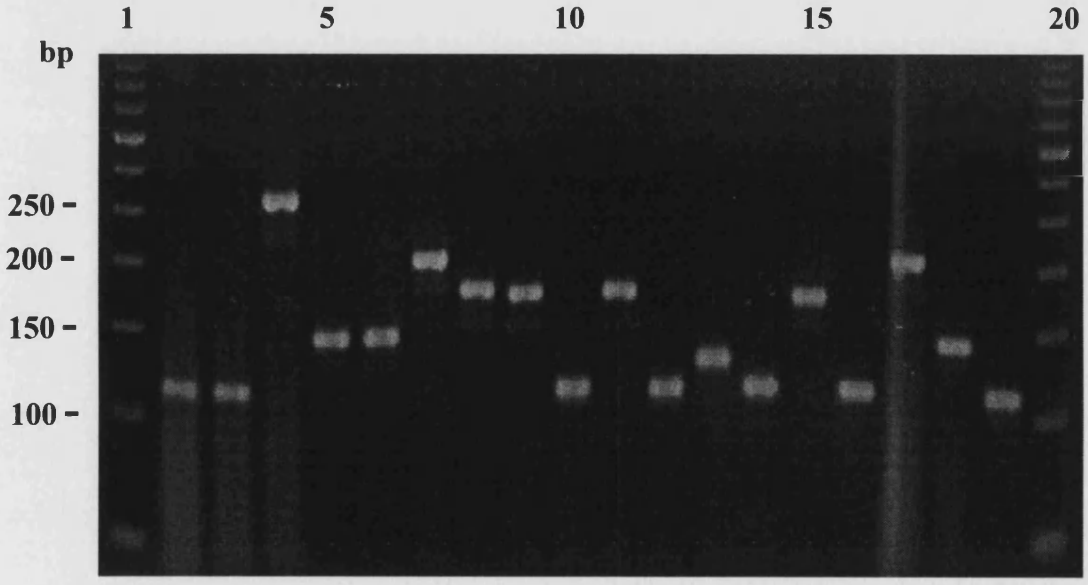



**Figure 5.14** PCR screen of 18 transformants resulting from a PCR amplification of ligation 1 (N-CAPC, TGC4, TGT4, C-CAPC) being cloned into pCR3. Lanes 1 and 20: 50 bp DNA ladder. Lanes 2-19: Transformants 1-18.

Using the second cloning strategy, the products of ligation 1 were digested with *Hind* III / *Xba* I (Figure 5.13, lane 3) and the DNA fragments sized between 50-150 bp were isolated (lanes 5 and 7) and cloned into pCR3. This range of *Hind* III / *Xba* I digested DNA fragments contains constructs I-IV (I:58bp, II:88bp, III:118bp, IV:148bp). Eighteen transformants were PCR screened (Figure 5.15) giving bands corresponding to constructs I-IV (I:7, II:3, III:5, IV:3), again with a bias towards construct I. Higher types of construct had been successfully excluded. Twelve inserts (3 of each construct type) were sequenced (Figure 5.16). All twelve inserts were "legal" genes confirming the fidelity of the assembly process and the success of the cloning strategy.

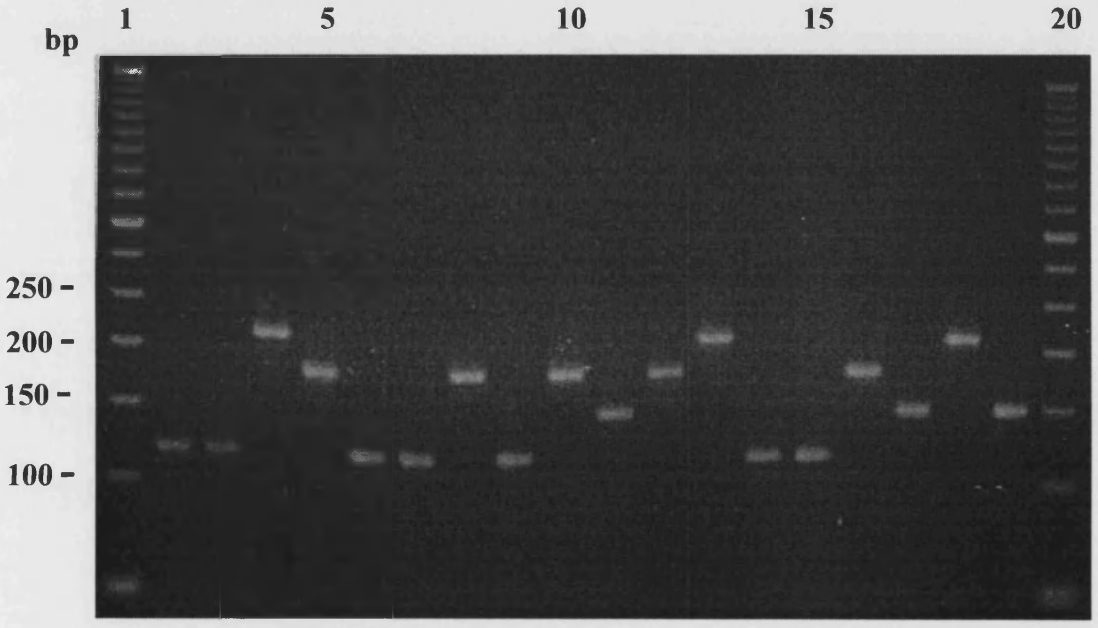



**Figure 5.15** PCR screen of 18 transformants resulting from a restriction digested ligation 1 (N-CAPC, TGC4, TGT4, C-CAPC) being cloned into pCR3. Lanes 1 and 20: 50 bp DNA ladder. Lanes 2-19: Transformants G1-18.

| Insert | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **G1** | N-CAPC | TGC4 | C-CAPC | | | | | | |
| **G2** | N-CAPC | TGC4 | C-CAPC | | | | | | |
| **G5** | N-CAPC | TGC4 | C-CAPC | | | | | | |
| **G10** | N-CAPC | TGC4 | TGT4 | TGC4 | C-CAPC | | | | |
| **G16** | N-CAPC | TGC4 | TGT4 | TGC4 | C-CAPC | | | | |
| **G18** | N-CAPC | TGC4 | TGT4 | TGC4 | C-CAPC | | | | |
| **G7** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC | | |
| **G9** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC | | |
| **G11** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC | | |
| **G3** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC |
| **G12** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC |
| **G17** | N-CAPC | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | TGT4 | TGC4 | C-CAPC |

**Figure 5.16** Block composition, derived from the nucleotide sequences, of twelve genes assembled by ligation 1 (N-CAPC, TGC4, TGT4, C-CAPC) and cloned by direct ligation into pCR3.

This experiment determined that for simple ligation reactions (containing relatively few species of block) the two cloning strategies produce a similar distribution of constructs I-IV in the recombinant DNA constructs, with a bias towards construct I.

### 5.2.2 Determination of the optimum cloning strategy for multi-block ligations

The aim of this experiment (Section 5.1.2) was to determine the optimum cloning strategy for multi-block ligation reactions. As the first four types of construct assembled by ligation 2 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) overlap in size (I:74-86bp, II:92-128bp, III:110-170bp, IV:128-212bp) the proportion of construct type in the recombinant DNA constructs cannot be determined by PCR screening (Construct I:109-121bp, II:127-163bp, III:145-205bp, IV:163-247bp). However, the two cloning strategies can be compared by looking for good variation in the sizes of the PCR products, and a low proportion of construct I.

Amplification of ligation 2 with primers specific for the terminating blocks gave a smear (Figure 5.17, lane 2) which was cloned into pCR3. Eighteen transformants

were PCR screened (Figure 5.18), 12 of which gave bands corresponding to construct I. Although the exact proportion of the other construct types in the recombinant DNA constructs could not be determined, five transformants produced bands sized between 140-170bp, representing construct types II, III or IV (Figure 5.1). One band, which was approximately 260bp in size, probably represented a very high construct type (i.e. V, VI, VII or VIII). In conclusion, this cloning strategy seems to preferentially clone type I constructs from multi-block ligation reactions.

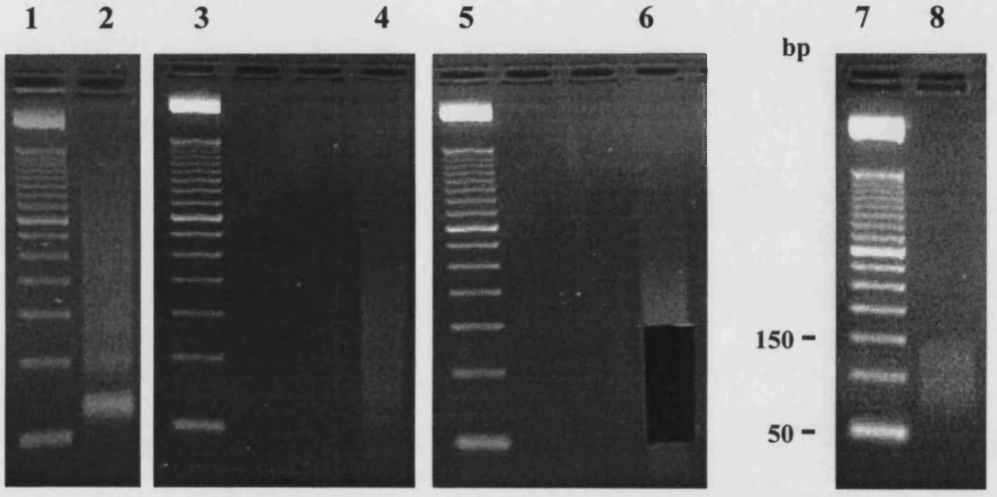



**Figure 5.17** The cloning of ligation 2 (N-CAPC, TGC2-6, TGT2-6, C-CAPC). Lane 1: 50 bp DNA ladder. Lane 2: PCR amplification of ligation 2. Lane 3: 50 bp DNA ladder. Lane 4: *Hind* III / *Xba* I digested products of ligation 2. Lane 5: 50 bp DNA ladder. Lane 6: Excision of 44-150bp DNA fragments from lane 4. Lane 7: 50 bp DNA ladder. Lane 8: Isolation of 44-150 bp DNA fragments from lane 4.

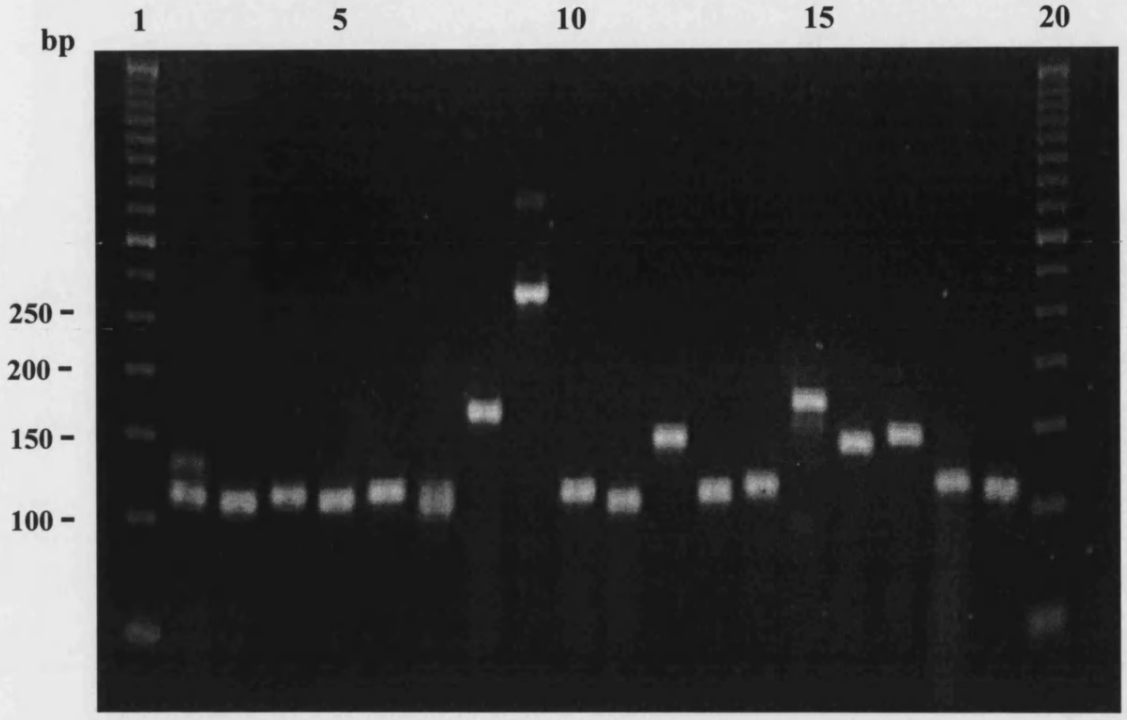



**Figure 5.18** PCR screen of 18 transformants resulting from a PCR amplification of ligation 2 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) being cloned into pCR3. Lanes 1 and 20: 50 bp DNA ladder. Lanes 2-19: Transformants 1-18.

Using the second cloning strategy, the products of ligation 2 were digested with *Hin*d III / *Xba* I (Figure 5.17, lane 4) and the DNA fragments sized between 44-150 bp were isolated (lanes 6 and 8) and cloned into pCR3. This range of DNA fragments excludes genes encoding peptides larger than 39 amino acids (a 149 bp DNA fragment encodes a peptide of 39 amino acids) and the vast majority of the genes encoding more than 8 cysteines, although some genes encoding 10 and 12 cysteines (constructs V and VI) will be present (Figure 5.1). Twenty-four transformants were PCR screened (Figure 5.19), 5 of which gave bands corresponding to construct I. Sixteen transformants produced bands sized between 140-215 bp, representing constructs II, III, IV or higher (Figure 5.1). Theoretically, the 44-150 bp range of *Hin*d III / *Xba* I digested DNA fragments, once cloned into pCR3 and PCR screened, would produce bands sized between 115-215 bp. However, two transformants produced bands larger than the theoretical maximum. One insert, which produced a band of approximately 260 bp (Figure 5.19, lane 2) was sequenced (Figure 5.20). This unexpectedly large

construct was a "legal" gene encoding 12 cysteines. This "oversized" construct revealed a limitation of the cloning strategy, as it had been formed from two incomplete genes, both of which had been isolated as part of the 44-150 bp range, ligating together because both were missing a terminating block.

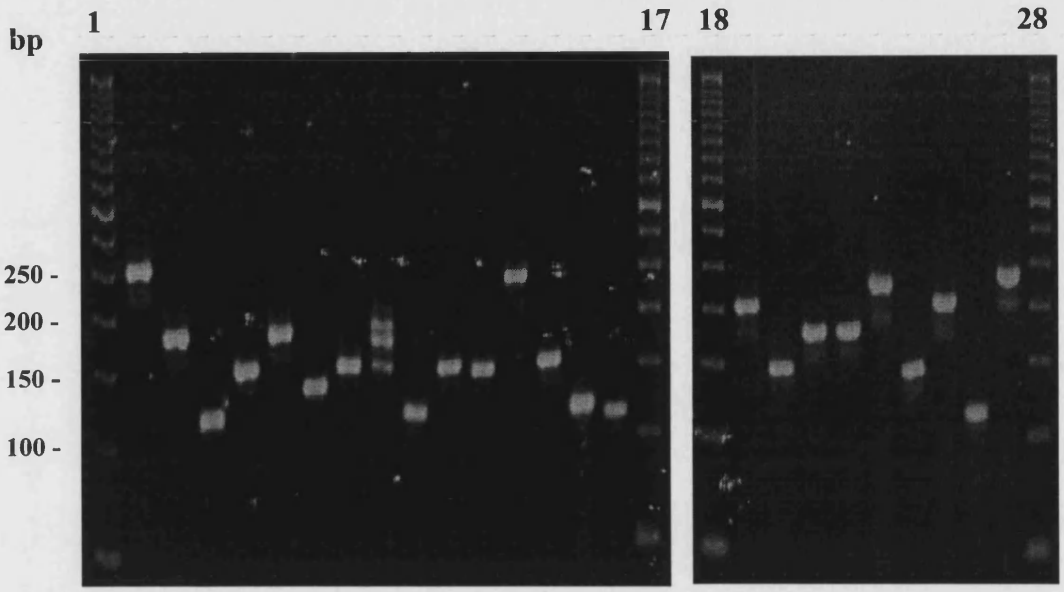



**Figure 5.19** PCR screen of 24 transformants resulting from a restriction digested ligation 2 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) being cloned into pCR3. Lanes 1, 17, 18 and 28: 50 bp DNA ladder. Lanes 2-16 and 19-27: Transformants H1-24, respectively.

| Insert | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| H1 | N-CAPC | TGC3 | TGT6 | TGC3 | TGT6 | TGC3 | TGT2 | TGC3 | TGT3 |
| | TGC3 | TGT3 | TGC5 | C-CAPC | | | | | |

**Figure 5.20** Block composition, derived from the nucleotide sequence, of a single "oversized" gene assembled by ligation 2 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) which should have been excluded during the cloning procedure.

This experiment determined that the optimum cloning strategy for multi-block ligation reactions is to ligate the assembled genes directly into pCR3 after digestion with *Hin*d III / *Xba* I. This cloning strategy produces a relatively even distribution of

constructs I-IV in the recombinant DNA constructs, and enables the assembled genes of interest to be selectively isolated.

### 5.2.3 Preliminary construction of a DNA library encoding a population of constrained peptides

This experiment (Section 5.1.3) served as a pilot experiment in determining the ability of multi-block ligation reactions to construct DNA libraries encoding populations of constrained peptides. The objective was to construct a simple DNA library using twelve species of block representing only the N, TGT, TGC and C classes.

The products of ligation 3 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) were digested with *Hin*d III / *Xba* I (Figure 5.21, lane 1) and the DNA fragments sized between 62-122 bp were isolated (lanes 4 and 8) and cloned into pCR3. This range of DNA fragments excludes genes encoding peptides larger than 30 amino acids (a 122 bp DNA fragment encodes a peptide of 30 amino acids), type I constructs (encoding 2 cysteines), and the vast majority of genes encoding more than 8 cysteines, although some genes encoding 10 cysteines (type V constructs) will be present (Figure 5.1). It was decided to exclude the type I constructs in light of the slight bias of the cloning strategy towards construct I (Sections 5.2.1 and 5.2.2) and as only five different type I constructs can be assembled by ligation 3, this precaution prevents the massive over-representation of the type I constructs in the DNA library.

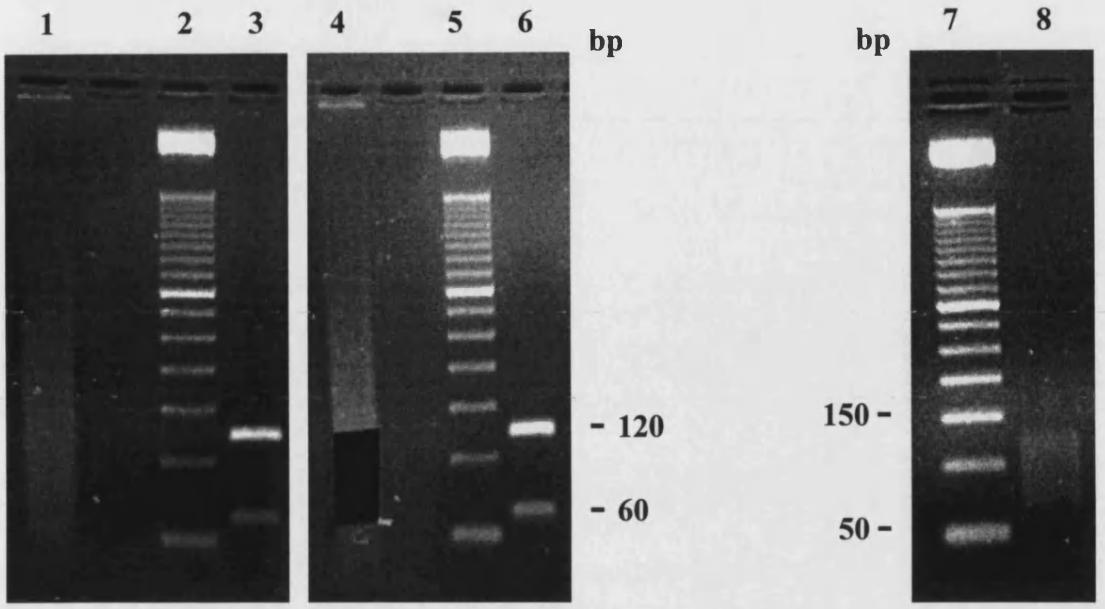



**Figure 5.21** The cloning of ligation 3 (N-CAPC, TGC2-6, TGT2-6, C-CAPC). Lane 1: *Hin*d III / *Xba* I digested products of ligation 3. Lanes 2 and 5: 50 bp DNA ladder. Lanes 3 and 6: 60 bp and 120 bp DNA markers. Lane 4: Excision of 62-122 bp DNA fragments from lane 1. Lane 7: 50 bp DNA ladder. Lane 8: Isolation of 62-122 bp DNA fragments from lane 1.

Eighteen transformants were PCR screened (Figure 5.22). Theoretically, the 62-122 bp range of *Hin*d III / *Xba* I digested DNA fragments, once cloned into pCR3 and PCR screened, would produce bands sized between 127-187 bp. The eighteen transformants produced bands within the upper size limit, although the type I constructs had not been successfully excluded, as five bands were between 109-121 bp in size. The other thirteen transformants produced bands sized between 130-175 bp, representing constructs II, III and IV. The variety of band sizes was suggestive of good diversity in the assembled genes.

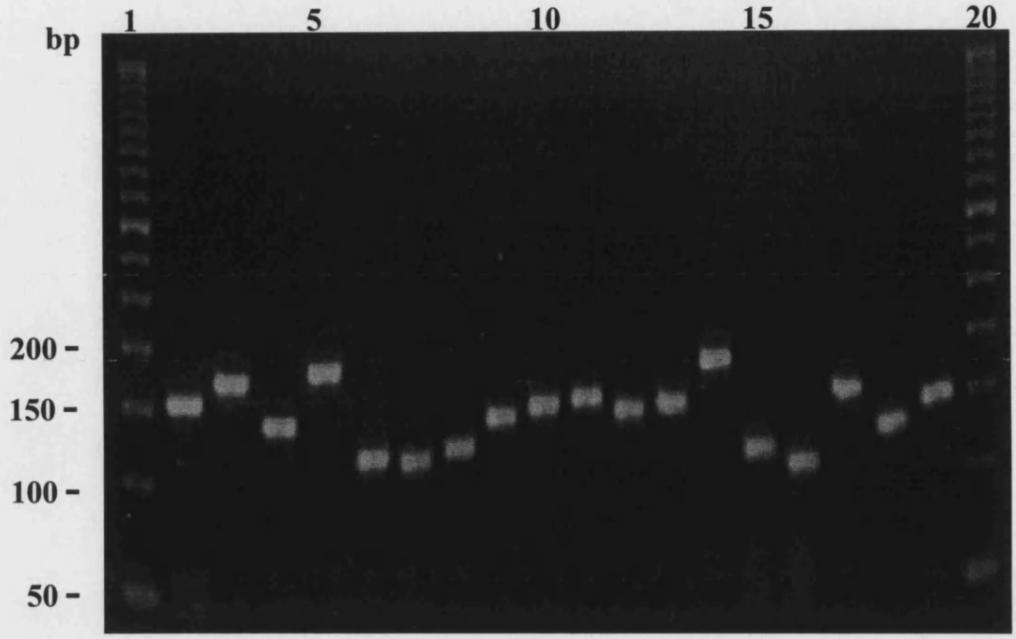



**Figure 5.22** PCR screen of 18 transformants resulting from a restriction digested ligation 3 (N-CAPC, TGC2-6, TGT2-6, C-CAPC) being cloned into pCR3. Lanes 1 and 20: 50 bp DNA ladder. Lanes 2-19: Transformants J1-18.

Five inserts, chosen because they were not type I constructs, were sequenced (Figure 5.23). All five inserts were "legal" genes (Constructs II:3 and III:2) with the only blocks not appearing in the sequenced inserts being TGC2 and TGT5. The sizes of the encoded peptides ranged from 14 to 27 amino acids, well within the specified limit of 30 amino acids. The results were encouraging, and suggested that multi-block ligation reactions are able to construct DNA libraries encoding populations of constrained peptides. Consequently, it was decided to construct more complicated DNA libraries (i.e. by including the TGC[ ] and TGT[ ] block classes in the multi-block ligations) and to sequence many more assembled genes.

| Insert | | | | | | Encoded Peptide (Amino Acids) | Construct Type |
|---|---|---|---|---|---|---|---|
| J1 | TGC4 | TGT6 | TGC5 | | | 19 | II |
| J2 | TGC3 | TGT4 | TGC3 | TGT3 | TGC5 | 24 | III |
| J3 | TGC4 | TGT2 | TGC4 | | | 14 | II |
| J4 | TGC3 | TGT2 | TGC6 | TGT6 | TGC4 | 27 | III |
| J8 | TGC4 | TGT2 | TGC4 | | | 14 | II |

**Figure 5.23** Propagating block composition, derived from the nucleotide sequences, of five genes assembled by ligation 3 (N-CAPC, TGC2-6, TGT2-6, C-CAPC). The length of the encoded peptide and construct type of the assembled gene is also detailed.

### 5.2.4 Construction of a DNA library encoding a population of constrained peptides: Part I

The objective of this experiment (Section 5.1.4) was to construct a DNA library encoding a population of constrained peptides, using fourteen species of block representing all six classes (N, TGC, TGT, TGC[ ], TGT[ ] and C).

The products of ligation 4 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC) were digested with *Hin*d III / *Xba* I (Figure 5.24, lane 1) and the DNA fragments sized between 62-122 bp were isolated (lanes 4 and 8) and cloned into pCR3. This range of DNA fragments excludes the genes encoding peptides larger than 30 amino acids (a 122 bp DNA fragment encodes a peptide of 30 amino acids), type I constructs (encoding 2 cysteines), and the vast majority of the genes encoding more than 8 cysteines, although some genes encoding 10 cysteines (type V constructs) will be present (Figure 5.1). The type I genes were excluded for the reasons detailed in Section 5.2.3.

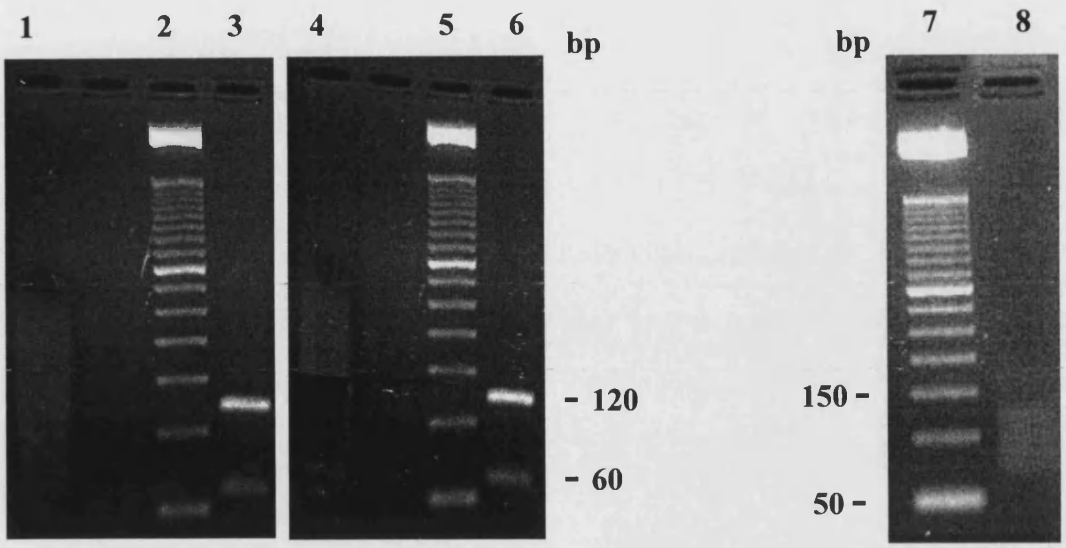



**Figure 5.24** The cloning of ligation 4 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). Lane 1: *Hin*d III / *Xba* I digested products of ligation 4. Lanes 2, 5 and 7: 50 bp DNA ladder. Lanes 3 and 6: 60 bp and 120 bp DNA markers. Lane 4: Excision of 62-122 bp DNA fragments from lane 1. Lane 8: Isolation of 62-122 bp DNA fragments from lane 1.

Twenty transformants were PCR screened (Figure 5.25). Theoretically, the 62-122 bp range of *Hin*d III / *Xba* I digested DNA fragments, once cloned into pCR3 and PCR screened, would produce bands sized between 127-187 bp. All the transformants produced bands within these size limits. All 20 inserts were sequenced (Figure 5.26). The inserts showed good gene diversity, with the sizes of the encoded peptides ranging from 13 to 24 amino acids, well within the specified maximum limit of 30 amino acids. Eighteen inserts were "legal" genes (Constructs II:16 and III:2). However, two inserts were "illegal" genes that had been assembled because the block overhangs had disobeyed standard base-pairing. The type I constructs had been successfully excluded. However, no type IV constructs (encoding 8 cysteines) had been cloned, this was probably due to the low proportion of type IV constructs present in the 62-122 bp range of *Hin*d III / *Xba* I digested DNA fragments that had been cloned into pCR3 (Figure 5.1).

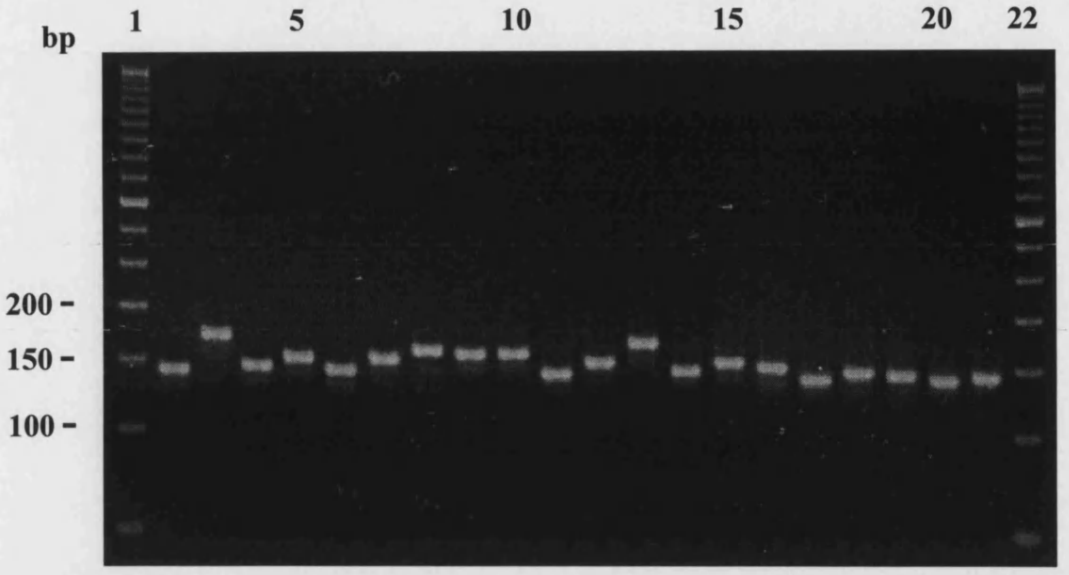



**Figure 5.25** PCR screen of 20 transformants resulting from a restriction digested ligation 4 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC) being cloned into pCR3. Lanes 1 and 22: 50 bp DNA ladder. Lanes 2-21: Transformants K1-20.

| Insert | | | | | | Encoded Peptide (Amino Acids) | Construct Type |
|---|---|---|---|---|---|---|---|
| K1 | TGC6 | TGC5 | | | | 14 | Illegal |
| K2 | TGC[CC]5 | TGC6 | TGT4 | TGC3 | | 24 | III |
| K3 | TGC4 | TGT3 | TGC4 | | | 15 | II |
| K4 | TGC4 | TGT4 | TGC6 | | | 18 | II |
| K5 | TGC3 | TGT3 | TGC5 | | | 15 | II |
| K6 | TGC4 | TGT4 | TGC6 | | | 18 | II |
| K7 | TGC6 | TGT6 | TGC5 | | | 21 | II |
| K8 | TGC6 | TGT5 | TGC5 | | | 20 | II |
| K9 | TGC6 | TGT5 | TGC5 | | | 20 | II |
| K10 | TGC[CC]5 | TGC5 | | | | 14 | II |
| K11 | TGC4 | TGT6 | TGC3 | | | 17 | II |
| K12 | TGC3 | TGT2 | TGC6 | TGT3 | TGC3 | 23 | III |
| K13 | TGC6 | TGT[CC]5 | | | | 15 | II |
| K14 | TGC4 | TGT6 | TGC3 | | | 17 | II |
| K15 | TGC5 | TGT3 | TGC4 | | | 16 | II |
| K16 | TGC4 | TGT[CC]5 | | | | 13 | II |
| K17 | TGC[CC]5 | TGT[CC]5 | | | | 15 | Illegal |
| K18 | TGC6 | TGT2 | TGC3 | | | 15 | II |
| K19 | TGC5 | TGT[CC]5 | | | | 14 | II |
| K20 | TGC6 | TGT[CC]5 | | | | 15 | II |

**Figure 5.26** Propagating block composition, derived from the nucleotide sequences, of twenty genes assembled by ligation 4 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). The length of the encoded peptide and construct type of the assembled gene is also detailed.

The overall incorporation frequencies of the twelve propagating blocks in the 20 sequenced inserts is very good (Figure 5.27), basically following the trend imposed upon ligation 4 at the outset (Section 5.1.4), with one significant exception which is the complete absence of TGC2. This problem is discussed in the summary.

| Block Class | Block Species | Incorporation Frequency |
|---|---|---|
| TGT | TGT2 | 2 |
| | TGT3 | 4 |
| | TGT4 | 3 |
| | TGT5 | 2 |
| | TGT6 | 3 |
| TGC | TGC2 | 0 |
| | TGC3 | 7 |
| | TGC4 | 8 |
| | TGC5 | 8 |
| | TGC6 | 11 |
| TGT[ ] | TGT[CC]5 | 5 |
| TGC[ ] | TGC[CC]5 | 3 |
| | | Total = 56 |

**Figure 5.27** The incorporation frequencies of the propagating blocks in the 20 sequenced genes assembled by ligation 4 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). Note the complete absence of TGC2.

Another anomaly of the sequencing results is that two pairs of inserts are identical. Inserts K8 and K9 (Figure 5.26) are identical type II constructs. The probability of these two constructs being identical by chance is 1:125 (as $5^3$ different type II constructs can be assembled with 5 species of TGT and 5 species of TGC block). It is theoretically possible that the two inserts independently assembled the same gene, although it is statistically unlikely, considering that only 20 inserts were sequenced. It is more likely that inserts K8 and K9 were actually the same recombinant DNA construct, as the two colonies that were screened from the transformation plate were 2.5mm apart. The same is also true for inserts K11 and K14 (Figure 5.26) which are identical type II constructs, whose colonies were 4mm apart on the transformation plate. However, other colonies had been picked from the transformation plates in a similar way, and their genes were not identical. For example, transformants K5 and K6 were 3mm apart on the transformation plates, and their inserts were different.

This anomaly casts doubt upon the uniqueness of a small proportion of inserts sequenced from previous ligation experiments. In any case, in the next experiment the screened transformants were a minimum of 1cm apart on the transformation plates.

### 5.2.5 Construction of a DNA library encoding a population of constrained peptides: Part II

This experiment (Section 5.1.5) was almost identical to the previous experiment (Section 5.1.4), the only difference being that the concentrations of the different sized TGT and TGC blocks in the ligation reaction had been staggered to increase the incidence of the smaller blocks in the DNA library. The size range of the assembled genes that were subsequently cloned, had also been increased in an attempt to clone a higher proportion of constructs III and IV.

The products of ligation 5 (N-CAPC, TGC2-6, TGT2-6, TGT[CC]5, TGC[CC]5, C-CAPC) were digested with *Hin*d III / *Xba* I (Figure 5.28, lane 3) and the DNA fragments sized between 62-150 bp were isolated (lanes 6 and 8) and cloned into pCR3. This range of DNA fragments excludes the genes encoding peptides larger than 39 amino acids (a 149 bp DNA fragment encodes a peptide of 39 amino acids), type I constructs (encoding 2 cysteines), and the vast majority of the genes encoding more than 8 cysteines, although some genes encoding 10 and 12 cysteines (constructs V and VI) will be present (Figure 5.1). The type I genes were excluded for the reasons detailed in Section 5.2.3. The size range was increased to 150 bp in order to clone some type IV constructs which contain more propagating blocks than the lower construct types, and which consequently provide more information on sequencing with respect to the frequency of block incorporation and the diversity of the assembled genes.

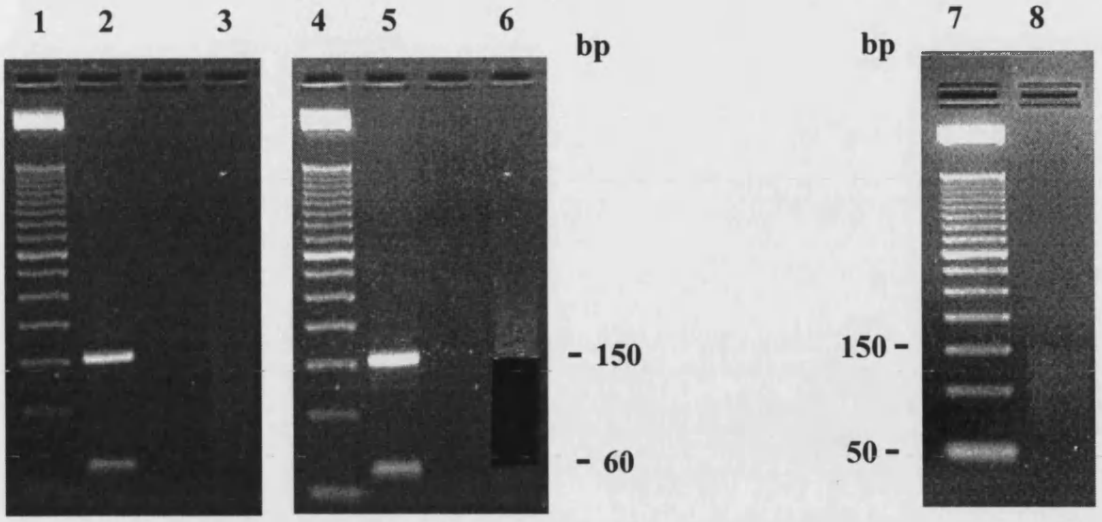



**Figure 5.28** The cloning of ligation 5 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). Lanes 1, 4 and 7: 50 bp DNA ladder. Lanes 2 and 5: 60 bp and 150 bp DNA markers. Lane 3: *Hin*d III / *Xba* I digested products of ligation 5. Lane 6: Excision of 62-150 bp DNA fragments from lane 3. Lane 8: Isolation of 62-150 bp DNA fragments from lane 3.

Eighteen transformants were PCR screened (Figure 5.29). Theoretically, the 62-150 bp range of *Hin*d III / *Xba* I digested DNA fragments, once cloned into pCR3 and PCR screened, would produce bands sized between 127-215 bp. All the transformants produced bands between approximately 120-210 bp in size. All 18 inserts were sequenced (Figure 5.30). The inserts showed good gene diversity, with the sizes of the encoded peptides ranging from 6 to 36 amino acids, within the specified maximum limit of 39 amino acids. Seventeen inserts were "legal" genes (Constructs I:1, II:7, III:4, IV:5). However, one insert was an "illegal" gene that had been assembled because the block overhangs had disobeyed standard base-pairing. The type I constructs had not been successfully excluded, although the higher types of construct (i.e. constructs V and VI) had been.

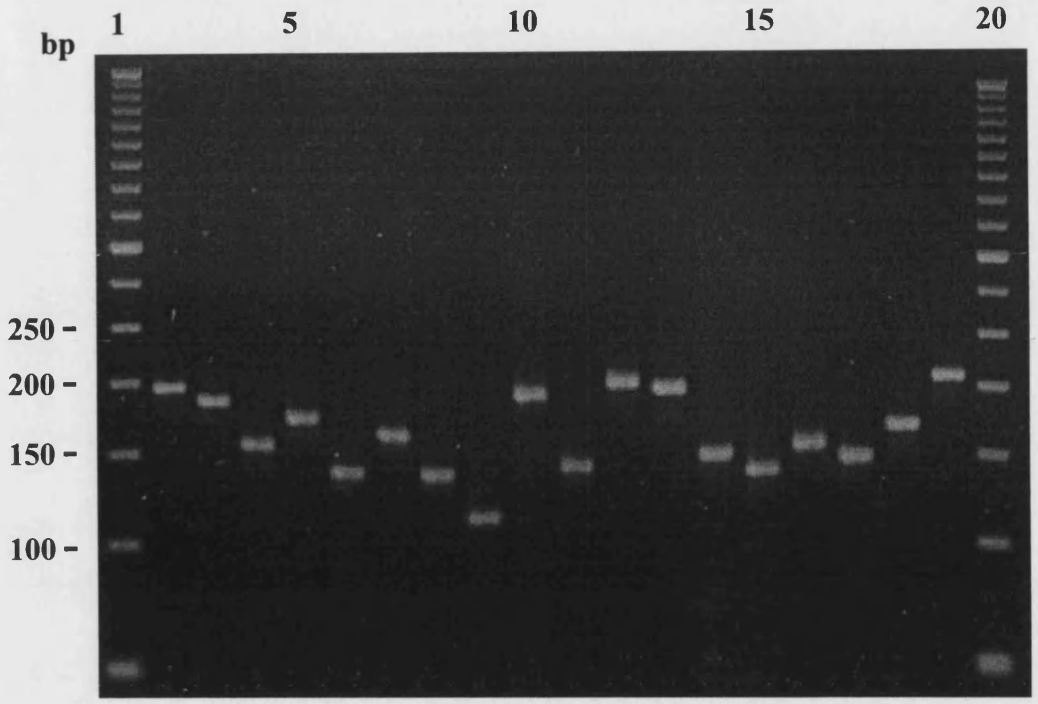



**Figure 5.29** PCR screen of 18 transformants resulting from a restriction digested ligation 5 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC) being cloned into pCR3. Lanes 1 and 20: 50 bp DNA ladder. Lanes 2-19: Transformants L1-18.

| Insert | | | | | | | | Encoded Peptide (Amino Acids) | Construct Type |
|---|---|---|---|---|---|---|---|---|---|
| L1 | TGC6 | TGT4 | TGC5 | TGT6 | TGC5 | | | 32 | III |
| L2 | TGC3 | TGT3 | TGC3 | TGT[CC]5 | TGT2 | TGC5 | | 29 | IV |
| L3 | TGC4 | TGT6 | TGC5 | | | | | 19 | II |
| L4 | TGC6 | TGT4 | TGC4 | TGT[CC]5 | | | | 25 | III |
| L5 | TGC[CC]5 | TGC5 | | | | | | 14 | II |
| L6 | TGC6 | TGT[CC]5 | TGT[CC]5 | | | | | 22 | III |
| L7 | TGC[CC]5 | TGC5 | | | | | | 14 | II |
| L8 | TGC4 | | | | | | | 6 | I |
| L9 | TGC3 | TGT3 | TGC3 | TGT[CC]5 | TGT6 | TGC4 | | 32 | IV |
| L10 | TGC3 | TGT5 | TGC3 | | | | | 15 | II |
| L11 | TGC6 | TGT3 | TGC4 | TGT[CC]5 | TGT5 | TGC4 | | 35 | IV |
| L12 | TGC6 | TGT4 | TGC3 | TGT[CC]5 | TGT4 | TGC3 | | 33 | IV |
| L13 | TGC4 | TGT4 | TGC5 | | | | | 17 | II |
| L14 | TGC3 | TGT3 | TGC4 | | | | | 14 | II |
| L15 | TGC6 | TGT3 | TGC6 | | | | | 19 | II |
| L16 | TGC3 | TGC3 | TGT[CC]5 | | | | | 16 | Illegal |
| L17 | TGC4 | TGT4 | TGC4 | TGT[CC]5 | | | | 23 | III |
| L18 | TGC6 | TGT3 | TGC3 | TGT3 | TGC6 | TGT3 | TGC4 | 36 | IV |

**Figure 5.30** Propagating block composition, derived from the nucleotide sequences, of eighteen genes assembled by ligation 5 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). The length of the encoded peptide and construct type of the assembled gene is also detailed.

The incorporation frequencies of the twelve propagating blocks in the 18 sequenced inserts (Figure 5.31) basically follow the trend imposed upon ligation 5 at the outset (Section 5.1.5). As in Section 5.2.4, the only real cause for concern is the complete absence of the TGC2 block and the very infrequent appearance of the TGT2 block.

| Block Class | Block Species | Incorporation Frequency |
|---|---|---|
| TGT | TGT2 | 1 |
| | TGT3 | 8 |
| | TGT4 | 6 |
| | TGT5 | 2 |
| | TGT6 | 3 |
| TGC | TGC2 | 0 |
| | TGC3 | 12 |
| | TGC4 | 11 |
| | TGC5 | 7 |
| | TGC6 | 9 |
| TGT[ ] | TGT[CC]5 | 9 |
| TGC[ ] | TGC[CC]5 | 2 |
| | | Total = 70 |

**Figure 5.31** The incorporation frequencies of the propagating blocks in the 18 sequenced genes assembled by ligation 5 (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC). Note the complete absence of TGC2 and the low prevalence of TGT2.

## 5.3 SUMMARY

This chapter determined that the optimum cloning strategy for genes assembled by multi-block ligation reactions is direct ligation into the cloning vector after restriction digestion. Crucially, this strategy produces a relatively even proportion of constructs I, II, III ... in the recombinant DNA constructs so that genes encoding 2, 4, 6 ... cysteines are equally represented in the DNA library, and enables the imposition of physical size limits upon the genes to be cloned. This allows the genes that we wish to express as the constrained peptide repertoire to be selectively isolated and subsequently cloned. Superfluous genes encoding peptides larger than 30 amino

acids, and the vast majority of the genes encoding more than 8 cysteines, can be successfully excluded.

The elucidation of the optimum cloning strategy had been the last stage in achieving the aim of this project. In order to test the whole procedure, two "incomplete" DNA libraries encoding populations of constrained peptides were constructed using fourteen species of block from the six block classes. The multi-block ligation reaction was the same in both experiments (N-CAPC, TGC2-6, TGT2-6, TGC[CC]5, TGT[CC]5, C-CAPC) although the concentrations of the blocks and the size ranges of the assembled genes which were subsequently cloned, varied slightly between the two experiments (Sections 5.1.4. and 5.1.5).

Of the 38 genes assembled by the two ligations which were sequenced (Figures 5.26 and 5.30) three were "illegal" genes which had been assembled because the block overhangs had disobeyed standard base-pairing (Section 4.3.2). However, since 92% of the assembled genes were "legal" genes encoding an even number of cysteine residues, and that these genes had been assembled because the six block classes had ligated together as intended, an 8% proportion of "illegal" genes in an assembled DNA library was deemed acceptable. Note that no "double-constructs" (Section 4.3.1) had been assembled, presumably because the *Xba* I enzyme that had been used throughout this chapter was from a brand new stock. The *Xba* I used throughout Chapter 4 (which had seen the assembly of many "double-constructs") was relatively old and consequently was probably low in activity.

The remaining 35 "legal" genes encoded an even number of cysteines, although construct II (encoding 4 cysteines) was the most common (I:1, II:23, III:6, IV:5). Both experiments had tried to exclude construct I, and the low proportion of constructs III and IV can be attributed to the size ranges of the assembled genes which were cloned into pCR3 in both experiments (Sections 5.1.4. and 5.1.5). Higher constructs (i.e. V, VI etc) had been successfully excluded as designed. The genes were encouragingly diverse, with respect to the encoded cysteine arrangements, the lengths of the encoded peptides (from 6 to 36 amino acids) and in the incorporation

frequencies of the different species of propagating block, with the exception of the TGC2 block which was completely absent, and the TGT2 block which had incorporated extremely infrequently. This problem with the TGC2 and TGT2 blocks had been totally unexpected as Sections 4.3.4 and 5.2.3 had unequivocally determined the ability of these blocks to incorporate and Section 4.3.7 had suggested that these blocks are not preferentially excluded during ligation reactions. Concatamers of the TGC[ ] and TGT[ ] supplementary blocks had only been observed in one gene. The sequences also revealed certain trends for the propagating blocks. The TGC[CC]5 block was only found ligated to N-CAPC (in contrast to its sister block TGT[CC]5 which incorporated between propagating blocks and to C-CAPC). However, TGC[CC]5 had only incorporated 5 times in the 38 genes and as the average number of propagating blocks in the sequenced inserts was low (see above) the chance of finding this block between propagating blocks was quite slim anyway.

In conclusion, this project has successfully conceived and tested a strategy, using six classes of sticky-ended dsDNA block, for constructing a DNA library encoding a population of variable-length peptides ( $\leq$ 30 amino acids in length) which contain an even number of randomly distributed cysteine residues (2, 4, 6 or 8) amongst fixed amino acids (Palmer *et al.*, 1998). Thus the aim of this project (Section 1.4) has been realised. Once constructed, such a DNA library should encode a structurally diverse repertoire of constrained peptides.

However, the 6 bp TGC2 and TGT2 blocks (which are the smallest blocks in the strategy) had unexpectedly not incorporated correctly into the genes assembled by the multi-block ligations. Although infuriating, this problem is not fatal to the strategy, as theoretically these blocks could be discarded and TGC[CXXC] and TGT[CXXC] (Figure 4.1) blocks could be used to cater for arrangements involving two cysteines separated by two amino acids. However, the consequence would be that the overall cost of constructing the DNA library would significantly increase, due to the extra number of oligonucleotides that would have to be synthesized to form these extra blocks. Fortunately, further investigation determined a means of incorporating the TGC2 and TGT2 blocks (Section 6.1), obviating the need for the extra blocks.

# CHAPTER 6

# PERSPECTIVE AND FUTURE WORK

## 6.1 INCORPORATION OF THE 6 bp TGT2 AND TGC2 BLOCKS

Subsequent research determined that in order for the 6 bp TGC2 and TGT2 blocks (the smallest of the strategy) to successfully incorporate into the genes assembled by multi-block ligation reactions, a number of criteria have to be fulfilled (M. R. Redfern, Department of Chemistry, University of Bath). The precise reasons why these criteria enable the 6 bp blocks to incorporate are unknown.

Firstly, by having a G:C base pair at either end of the spacer, the blocks are probably less susceptible to fraying and are therefore probably more stable. This requirement limits the first amino acid encoded by the spacer to just 10 (A, R, D, Q, E, G, H, L, P and V) whilst the identity of the second amino acid is unaffected. Secondly, the nucleotide sequences of the spacers in these blocks must be asymmetric rather than palindromic, so that the oligonucleotides comprising the blocks are unable to self-anneal. Thirdly, the amounts of these blocks in the multi-block ligation reactions must be increased to five times their theoretical value. For example, if 10pmol of each of the TGC3-6 blocks is present then 50pmol of TGC2 must be present for it to incorporate with the same frequency as these blocks. Two redesigned versions of TGC2 (Figure 6.1), denoted TGC2[*Bsr*B I] and TGC2[*Ear* I] after the restriction site each carries, both fulfill the first two requirements and consequently successfully incorporate during multi-block ligation reactions when present at five times their theoretical amount.

Presumably, the infrequent inclusion of TGT2 and the complete absence of TGC2 from the genes assembled by the multi-block ligation reactions in Sections 5.2.4 and

5.2.5, had been due to oligonucleotide self-annealing lowering the amount of the TGT2 block and instability of the TGC2 block.

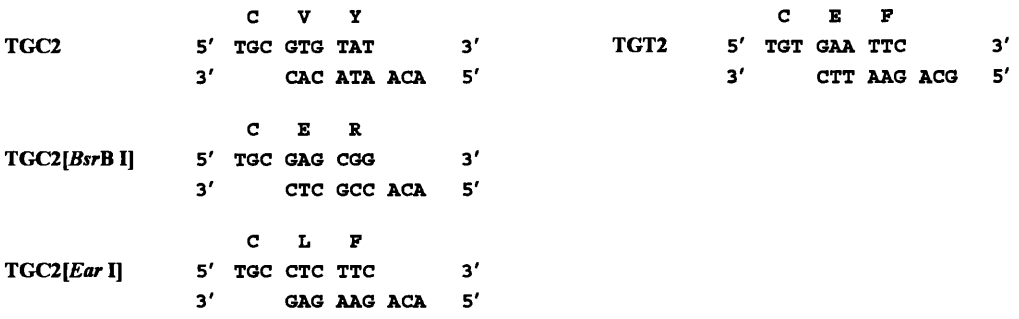



**Figure 6.1** The original TGT2 and TGC2 blocks alongside two redesigned versions of TGC2, denoted TGC2[*Bsr*B I] and TGC2[*Ear* I], which successfully incorporate during multi-block ligation reactions when present at five times their theoretical amount.

## 6.2 CONSTRUCTION OF THE DNA LIBRARY

The strategy for constructing the DNA library imposes two significant limitations upon the encoded peptides: the minimum and maximum inter-cysteine distance is fixed at the outset by the design of the dsDNA blocks, and the maximum number of contiguous cysteines in any peptide is three. The strategy is therefore unable to construct a DNA library encoding all possible cysteine arrangements. However, the inherent limitations are actually advantageous for our purposes since we wish the inter-cysteine distance to vary from 0-6 amino acids, and to minimise the proportion of peptides in the repertoire containing more than two contiguous cysteines, roughly in keeping with the conotoxins (Section 1.1.2). Note that peptides containing three contiguous cysteines will represent an extremely small proportion of the repertoire and that one conotoxin (SII) does have three contiguous cysteines (Ramilo *et al.*, 1992).

The full complement of forty-six dsDNA propagating blocks needed to construct the "complete" DNA library is TGT[$X_{2\text{-}6}$], TGT[CC][$X_{1\text{-}6}$], TGT[CXC][$X_{1\text{-}6}$], TGT[$X_{1\text{-}6}$][CC], TGT[$X_{1\text{-}6}$][CXC], TGC[$X_{2\text{-}6}$], TGC[CC][$X_{1\text{-}6}$] and TGC[CXC][$X_{1\text{-}6}$] (Figure 6.2). Assuming that the amino acids beyond the terminal cysteines contribute little to the shape of the peptide, only a single species of each N and C block needs to be added to this mix. However, it would be possible to vary the number of amino acids beyond the terminal cysteines, if so desired, by encoding variable length spacers in the N and C blocks. The sequences of the N-CAPC and C-CAPC blocks will have to be redesigned so that the assembled DNA library can be cloned into a suitable vector, and expressed on the surface of filamentous bacteriophage by fusion to the pVIII coat protein (Section 6.3). The TGT[$X_{1\text{-}6}$][CC] and TGT[$X_{1\text{-}6}$][CXC] blocks encode [CC] and [CXC] at their "C-terminus", respectively, in contrast to the TGT[CC][$X_{1\text{-}6}$] and TGT[CXC][$X_{1\text{-}6}$] blocks (Figure 4.1) which encode [CC] and [CXC] at their "N-terminus", respectively.

## TGC[CC][$X_{1-6}$]

```
    C   C   G
5'  TGC TGC GGC      3'
3'      ACG CCG ACG  5'

    C   C   F   G
5'  TGC TGT TTC GGG      3'
3'      ACA AAG CCC ACG  5'

    C   C   V   N   Y
5'  TGC TGT GTT AAC TAT      3'
3'      ACA CAA TTG ATA ACG  5'

    C   C   D   A   S   G
5'  TGC TGT GAC GCC TCA GGC      3'
3'      ACA CTG CGG AGT CCG ACG  5'

    C   C   S   T   K   G   E
5'  TGC TGT TCT ACC AAA GGT GAA      3'
3'      ACA AGA TGG TTT CCA CTT ACG  5'

    C   C   S   A   P   G   T   L
5'  TGC TGT TCA GCC CCA GGA ACA CTT      3'
3'      ACA AGT CGG GGT CCT TGT GAA ACG  5'
```

## TGC[CXC][$X_{1-6}$]

```
    C   F   C   D
5'  TGC TTC TGC GAC      3'
3'      AAG ACG CTG ACG  5'

    C   I   C   K   M
5'  TGC ATA TGC AAG ATG      3'
3'      TAT ACG TTC TAC ACG  5'

    C   N   C   Q   R   S
5'  TGC AAC TGC CAG CGA AGC      3'
3'      TTG ACG GTC GCT TCG ACG  5'

    C   G   C   Y   W   P   V
5'  TGC GGT TGC TAC TGG CCA GTC      3'
3'      CCA ACG ATG ACC GGT CAG ACG  5'

    C   A   C   P   H   K   Q   R
5'  TGC GCA TGC CCG CAT AAG CAG CGT      3'
3'      CGT ACG GGC GTA TTC GTC GCA ACG  5'

    C   T   C   S   A   R   G   D   Q
5'  TGC ACA TGC AGT GCA AGG GGC GAT CAG      3'
3'      TGT ACG TCA CGT TCC CCG CTA GTC ACG  5'
```

## TGT[CC][$X_{1-6}$]

```
    C   C   N
5'  TGT TGC AAC      3'
3'      ACG TTG ACA  5'

    C   C   P   G
5'  TGT TGT CCA GGC      3'
3'      ACA GGT CCG ACA  5'

    C   C   N   P   A
5'  TGT TGC AAT CCT GCT      3'
3'      ACG TTA GGA CGA ACA  5'

    C   C   D   L   I   Y
5'  TGT TGC GAT CTA ATC TAT      3'
3'      ACG CTA GAT TAG ATA ACA  5'

    C   C   S   T   K   G   E
5'  TGT TGC TCT ACC AAA GGT GAA      3'
3'      ACG AGA TGG TTT CCA CTT ACA  5'

    C   C   T   S   W   F   P   R
5'  TGT TGC ACC TCC TGG TTC CCT AGA      3'
3'      ACG TGG AGG ACC AAG GGA TCT ACA  5'
```

## TGT[CXC][$X_{1-6}$]

```
    C   K   C   G
5'  TGT AAA TGC GGA      3'
3'      TTT ACG CCT ACA  5'

    C   G   C   S   P
5'  TGT GGG TGC AGC CCC      3'
3'      CCC ACG TCG GGG ACA  5'

    C   T   C   G   I   S
5'  TGT ACC TCC TGG TTC CCT      3'
3'      TGC ACG CCA TAG AGA ACA  5'

    C   S   C   V   R   S   Y
5'  TGT TCC TGC GTC CGA TCA TAC      3'
3'      AGG ACG CAG GCT AGT ATG ACA  5'

    C   N   C   P   Q   A   L   P
5'  TGT AAT TGC CCA CAG GCA CTG CCT      3'
3'      TTA ACG GGT GTC CGT GAC GGA ACA  5'

    C   Q   C   F   N   H   L   D   D
5'  TGT CAA TGC TTT AAC CAC TTA GAT GAC      3'
3'      GTT ACG AAA TTG GTG AAT CTA CTG ACA  5'
```

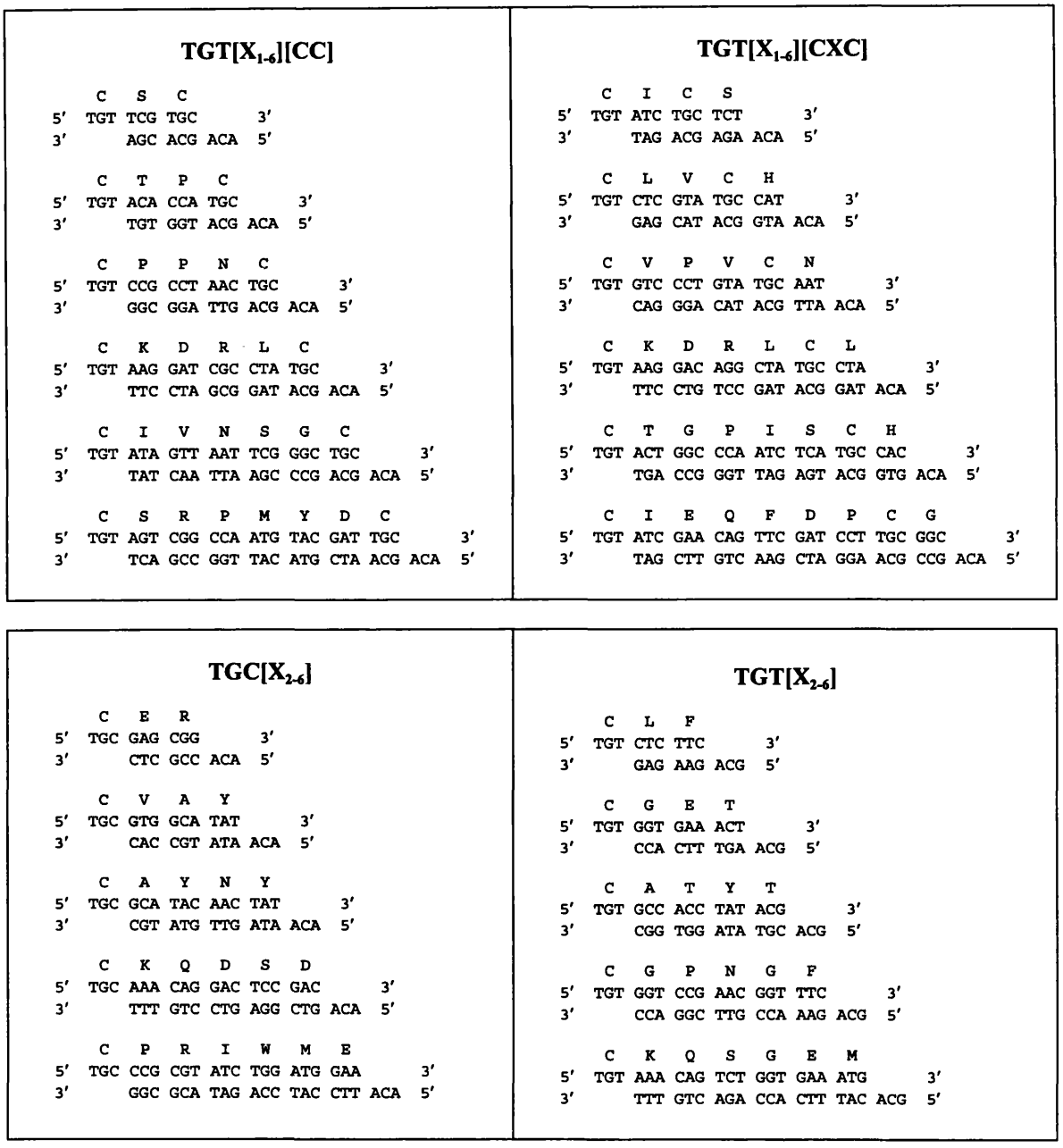



**Figure 6.2** The full complement of dsDNA propagating blocks needed to construct the DNA library. The constructed DNA library will encode a combinatorial peptide library in which length and cysteine arrangement varies, that should generate a structurally diverse repertoire of ~$10^7$ constrained peptides. Note that the inter-cysteine distance is limited to 0-6 amino acids, no more than three cysteines may be contiguous in any peptide (although triple cysteine arrangements would actually be relatively rare), and a small proportion of desirable cysteine arrangements are not possible (see text for details). The sequences encoded by TGC4, TGT2 and TGC2 have been redesigned (Figure 4.2).

The frequencies of individual amino acids in the inter-cysteine loops of the conotoxins are detailed (Figure 6.3). The amino acids encoded by the spacers in the dsDNA blocks (Figure 6.2) mirror this amino acid distribution (except for 4-*trans*-hydroxyproline which is not represented). No attempt was made to bias the codon usage in the dsDNA blocks to any particular species, since the DNA library may be expressed in a number of different organisms (e.g. *E. coli*, *Saccharomyces cerevisiae* etc.).

| Amino Acid | Frequency | % |
|---|---|---|
| A | 29 | 4 |
| D | 29 | 4 |
| E | 12 | 2 |
| F | 30 | 4 |
| G | 81 | 12 |
| H | 14 | 2 |
| I | 31 | 4 |
| K | 46 | 6 |
| L | 34 | 5 |
| M | 8 | 1 |
| N | 39 | 6 |
| O | 39 | 6 |
| P | 38 | 5 |
| Q | 25 | 4 |
| R | 55 | 8 |
| S | 79 | 11 |
| T | 33 | 5 |
| V | 31 | 4 |
| W | 9 | 1 |
| Y | 40 | 6 |
| **Total** | **702** | **100%** |

**Figure 6.3** The frequencies of individual amino acids in the inter-cysteine loops of the conotoxins. Note that a total of 702 amino acids were sampled. Data supplied by J. P. L. Cox, Department of Chemistry, University of Bath. The amino acids are designated according to the standard amino acid code (Sambrook *et al.*, 1989) with the exception of O which represents 4-*trans*-hydroxyproline.

The constructed DNA library should encode a structurally diverse repertoire of ~$10^7$ constrained peptides that are ≤ 30 amino acids in length and which contain either 2, 4, 6, or 8 cysteines amongst fixed amino acids[1] (G. C. Smith, Department of Mathematical Science, University of Bath). The number of constrained peptide folded main chain conformations is equal to the number of different cysteine arrangements, multiplied by the number of possible disulphide pairings for each arrangement (2, 4, 6 and 8 cysteines can pair in 1, 3 [3x1], 15 [5x3x1] and 105 [7x5x3x1] different ways, respectively). The mathematical calculation does not take into account any triple cysteine arrangements. Peptides larger than 30 amino acids in length, and the vast majority encoding more than 8 cysteines, can be excluded from the peptide repertoire at the cloning stage of the DNA library (Section 5.3). Note the similarities between the conotoxins (Section 1.1.2) and the peptide repertoire with respect to length, the even number of cysteine residues, the number of disulphide-bonds, the length and amino acid composition of the inter-cysteine loops, and the permissible cysteine arrangements. The constrained peptide repertoire should be a potentially rich source of novel peptide ligands that could serve as scaffolds for the presentation of further combinatorial peptide libraries.

If the block spacers encoded random amino acids, the size of the peptide repertoire would become astronomic, giving rise to ~$10^{33}$ different sequences, greatly exceeding the maximum size of conventional phage repertoires (Section 1.2.2). However, it must be noted that molecules that bind to a chosen target have been isolated from repertoires that have suffered from a similar lack of sequence space (Bock *et al.*, 1992). Random amino acids could also introduce cysteines into the spacers of the encoded peptides, resulting in some peptides carrying an odd number of cysteines, which may disrupt their folding. However, the grand strategy is to introduce sequence diversity after a peptide that is complementary in "shape" to the chosen target has

[1] [The number of possible cysteine arrangements in a sequence of length L containing N cysteines is the coefficient of $X^{L-N}$ in the polynomial $(1 + X + X^2 + \ldots + X^6)^{N-1}$ once it has been expanded.]

been selected (Section 6.3), thus fixed amino acids are to be encoded by the block spacers.

The limitations inherent to the strategy exclude a multiplicity of undesirable cysteine arrangements, considerably reducing the overall size of the DNA library. This is advantageous since a repertoire of ~$10^7$ different peptides is well within the maximum size of a conventional phage repertoire (Section 1.2.2). For example, with no limitations, eight cysteines can be arranged 5.85 x $10^6$ [30!/(8!22!)] ways in a 30 amino acid peptide, giving rise to 6.15 x $10^8$ [5852925 x 105] different folded main chain conformations, which alone exceeds the maximum size of a phage repertoire. Note that the quoted sizes of all the peptide repertoires represent a theoretical maximum, and that many peptide conformations would be excluded on stereochemical grounds whilst others would be disfavoured energetically.

The repertoire of constrained peptides has a number of limitations. Firstly, due to the poor sequence diversity (caused by using fixed amino acids in the block spacers) some peptides that are complementary in "shape" to the chosen target may actually be electrostatically incompatible, and consequently will not be selected. Another limitation concerns the folding of the peptides once they are expressed on the surface of filamentous bacteriophage by fusion to the pVIII coat protein (Section 1.2.2). A peptide containing 2, 4, 6 or 8 cysteines can theoretically form 1, 3, 15 and 105 possible disulphide pairings, respectively. Consequently, a unique cysteine arrangement encoded by a bacteriophage genome will probably be expressed on the surface of the bacteriophage as lots of different conformations rather than just one. This phenomenon could affect the success of the multivalent display (Section 1.2.2). It is interesting to note that conotoxins (Section 1.1.2) are expressed as 60-90 amino acid propeptides, which on folding are cleaved to release the mature conotoxin (Woodward *et al.*, 1990)(Olivera *et al.*, 1990a, 1995). Presumably, the propeptide directs the folding of each conotoxin into its biologically active disulphide framework. However, there is evidence to suggest that conotoxins are able to fold correctly without the propeptide, and that the biologically active disulphide framework is the most readily formed and most stable conformer, suggesting it is thermodynamically

favoured (Nishiuchi and Sakakibara, 1982)(Cruz *et al.*, 1989)(Pricecarter *et al.*, 1996). Finally, a small number of desirable cysteine arrangements cannot be created by the blocks (Figure 6.2), specifically those generated by the TGT[ ] and TGC[ ] block classes alone. Thus cysteine arrangements involving only two contiguous cysteines (e.g. CC...CC...CC), or two cysteines separated by a single amino acid (e.g. CXC...CXC...CXC), or mixtures of both (e.g. CXC...CC...CXC), will not be represented in the constructed DNA library.

## 6.3 GENERAL PERSPECTIVE OF THIS PROJECT

The work described in this project is part of a grand strategy for discovering novel pharmaceuticals for human therapy.

Once constructed, the DNA library encoding the structurally diverse repertoire of ~$10^7$ constrained peptides (Section 6.2) is to be expressed as a "multivalent" peptide-geneVIII phage library (Section 1.2.2). This approach facilitates the selection of peptides that are complementary in "shape" to the chosen target, but which bind with a relatively low affinity due to the lack of sequence diversity within the peptide repertoire. The structural diversity of the peptide repertoire should enable a "weakly-binding" peptide to be selected for any chosen target. Such a peptide repertoire would be analogous to the naïve primary antibody repertoire of the human immune system and a potentially rich source of pharmaceutical leads. Note that the peptide repertoire, which represents a multiplicity of constrained topologies, is effectively a "library of libraries" from which a peptide with a suitable constrained topology can be selected for any chosen target.

A second combinatorial peptide library is then presented from the selected constrained peptide scaffold in order to improve its binding affinity towards the target. Constrained peptides make excellent scaffolds for the presentation of combinatorial peptide libraries (Section 1.2.5). The gene encoding the selected constrained peptide can be subjected to point mutagenesis and genetic recombination to form a second

DNA library encoding a repertoire of conformationally homogeneous peptides with varying affinity for the target. Point mutagenesis can be introduced using either error-prone PCR (Bartel and Szostak, 1993) or oligonucleotide-directed mutagenesis (Oliphant *et al.*, 1986)(Bedwell *et al.*, 1989), whilst genetic recombination can be effected using either "sexual" PCR (Stemmer, 1994a, 1995a)(Smith, 1994) or exon shuffling (Fisch *et al.*, 1996).

Expression of the second DNA library as a "monovalent" peptide-geneIII phage library (Section 1.2.2) coupled with recursive cycles of point mutagenesis, genetic recombination, and selection on the basis of improved binding affinity towards the target, should evolve "strongly-binding" peptides via *in vitro* evolution (Stemmer, 1995b)(Brookfield, 1995). Stringent selection should ensure that sequence hypervariability is only introduced to the inter-cysteine loops of the constrained peptide scaffold and that its three-dimensional structure is unaltered. Note that this strategy for generating high-affinity peptides is akin to the strategy used by the human immune system to generate high-affinity antibodies (Section 1.1.1).

The efficacy of such a strategy was demonstrated when a 20 amino acid peptide mimic of the hormone erythropoietin was isolated (Wells, 1996)(Wrighton *et al.*, 1996)(Livnah *et al.*, 1996). However, in this work the initial peptide repertoire was a library of random cyclic peptides [$CX_8C$] (Section 1.2.3).

Peptides elicit important biological functions in the human body. For example, peptides comprise the largest class of hormones (Moore, 1994) and fulfill a myriad of other biological roles exemplified by the anti-microbial activity of the defensins (Harder *et al.*, 1997). Peptides isolated from natural sources have also been used therapeutically in humans. For example, conotoxins have been used as diagnostics (Olivera *et al.*, 1990a)(Sher *et al.*, 1989) and as analgesics (Law and Brose, 1997)(Guddat *et al.*, 1996). However, peptides do not generally make good pharmaceuticals *in vivo* (apart from as vaccines) as they are prone to enzymatic degradation, rapid clearance, are able to trigger an immune response, and suffer from poor pharmacological properties (Schumacher *et al.*, 1996).

One way to circumvent these limitations is to use mirror-image phage display to isolate D-peptides that have the desired activity towards a chosen target, but which are resistant to degradation by naturally occurring enzymes (Schumacher *et al.*, 1996). Another approach is to synthesize peptidomimetic compounds that mimic the essential pharmacological features of the selected bioactive peptides on a non-peptide scaffold (Adang *et al.*, 1994)(Etzkorn *et al.*, 1994)(Beeley, 1994)(Bianchi *et al.*, 1995)(Moore, 1994).

Thus the structurally diverse repertoire of constrained peptides represents the first and most significant link in a chain that could lead to the creation of novel pharmaceuticals for human therapy. The chain involves the isolation of peptides that have the desired pharmacological activity towards the therapeutic target ("the pharmaceutical leads") and the rational conversion of their activity into small-molecule mimetics suitable for human therapy. Ideal therapeutic targets in the field of cancer research could include molecules involved in angiogenesis (Bussolino *et al.*, 1997) and metastasis (Davidson *et al.*, 1997)(Gomis-Rüth *et al.*, 1997).

# REFERENCES


**Adang**, A.E.P., Hermkens, P.H.H., Linders, J.T.M., Ottenheijm, H.C.J. and Vanstaveren, C.J. (1994) "Case-histories of peptidomimetics – progression from peptides to drugs" Receuil des Travaux Chimiques des Pays-Bays – Journal of the Royal Netherlands Chemical Society 113, 2, 63-78.

**Akrigg**, A., Wilkinson, G.W.G., and Oram, J.D. (1985) "The structure of the major immediate early gene of human cytomegalovirus strain AD169" Virus Research 2, 107-121.

**Andersen**, J.S., Svensson, B. and Roepstorff, P. (1996) "Electrospray ionization and matrix assisted laser desorption/ionization mass spectrometry: Powerful analytical tools in recombinant protein chemistry" Nature Biotechnology 14, 449-457.

**Bai**, J., Liu, Y.H., Lubman, D.M. and Siemieniak, D. (1994) "Matrix-assisted laser-desorption ionization mass-spectrometry of restriction enzyme-digested plasmid DNA using an active Nafion substrate" Rapid Communications in Mass Spectrometry 8, 9, 687-691.

**Barbato**, G., Cicero, D.O., Bianchi, E., Pessi, A. and Bazzo, R. (1996) "High-resolution solution structure of 2 members of a conformationally homogeneous combinatorial peptide library based on the classical zinc-finger motif" Journal of Biomolecular NMR 8, 1, 36-48.

**Barhanin**, J., Schmid, A. and Lazdunski, M. (1988) "Properties of structure and interaction of the receptor for ω-conotoxin, a polypeptide active on $Ca^{2+}$ channels" Biochemical and Biophysical Research Communications 150, 2, 1051-1062.

**Bartel**, D.P. and Szostak, J.W. (1993) "Isolation of new ribozymes from a large pool of random sequences" Science 261, 1411-1418.

**Bedwell**, D.M., Strobel, S.A., Yun, K., Jongeward, G.D. and Emr, S.D. (1989) "Sequence and structural requirements of a mitochondrial protein import signal defined by saturation cassette mutagenesis" Molecular and Cellular Biology 9, 3, 1014-1025.

**Beeley**, N. (1994) "Peptidomimetics and small-molecule drug design – towards improved bioavailability and *in vivo* stability" Trends in Biotechnology 12, 6, 213-216.

**Berek**, C. and Milstein, C. (1987) "Mutation drift and repertoire shift in the maturation of the immune response" Immunological Reviews 96, 23-41.

**Better**, M., Chang, C., Robinson, R. and Horwitz, A. (1988) "*Escherichia coli* secretion of an active chimeric antibody fragment" Science 240, 1041-1043.

**Bianchi**, E., Folgori, A., Wallace, A., Nicotra, M., Acali, S., Phalipon, A., Barbato, G., Bazzo, R., Corteze, R., Felici, F. and Pessi, A. (1995) "A conformationally homogeneous combinatorial peptide library" J. Mol. Biol. 247, 154-160.

**Bock**, L.C., Griffin, L.C., Latham, J.A., Vermaas, E.H. and Toole, J.J. (1992) "Selection of single-stranded DNA molecules that bind and inhibit human thrombin" Nature 355, 564-566.

**Bonnycastle**, L.L.C., Mehroke, J.S., Rashed, M., Gong, X. and Scott, J.K. (1996) "Probing the basis of antibody reactivity with a panel of constrained peptide libraries displayed by filamentous phage" J. Mol. Biol. 258, 747-762.

**Borrebaeck**, C.A.K. (1995) "Antibody engineering" 2nd Edition, Oxford University Press, UK.

**Boshart**, M., Weber, F., Jahn, G., Dorsch-Häsler, K., Fleckenstein, B. and Schaffner, W. (1985) "A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus" Cell 41, 521-530.

**Brenner**, S. and Milstein, C. (1966) "Origin of antibody variation" Nature 221, 242-243.

**Broach**, J.R. and Thorner, J. (1996) "High-throughput screening for drug discovery" Nature 384, Supplement, 14-16.

**Brookfield**, J.F.Y. (1995) "Making selection work" Nature 375, 449.

**Bussolino**, F., Mantovani, A. and Persico, G. (1997) "Molecular mechanisms of blood vessel formation" Trends in Biochemical Sciences 22, 7, 251-256.

**Cheng**, X., Camp, D.G., Wu, Q., Bakhtiar, R., Springer, D.L., Morris, B.J., Bruce, J.E., Anderson, G.A., Edmonds, C.G. and Smith, R.D. "Molecular weight determination of plasmid DNA using electrospray ionization mass spectrometry" Nucleic Acids Research 24, 11, 2183-2189.

**Choo**, Y. and Klug, A. (1995) "Designing DNA-binding proteins on the surface of filamentous phage" Current Opinion in Biotechnology 6, 431-436.

**Chothia**, C. and Lesk, A.M. (1987) "Canonical structures for the hypervariable regions of immunoglobulins" J. Mol. Biol. 196, 901-917.

**Chothia**, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M. and Poljak, R.J. (1989) "Conformations of immunoglobulin hypervariable regions" Nature 342, 877-883.

**Clackson**, T. and Wells, J.A. (1994) "*In vitro* selection from protein and peptide libraries" Trends in Biotechnology 12, 173-184.

**Clark**, J.M. (1988) "Novel non-templated nucleotide addition reactions catalysed by prokaryotic and eukaryotic DNA polymerases" Nucleic Acids Research 16, 9677-9686.

**Cohen**, S.N., Chang, A.C.Y. and Hsu, L. (1972) "Nonchromosomal antibiotic resistance in bacteria: genetic transformation of *Escherichia coli* by R-factor DNA" Proc. Natl. Acad. Sci. USA. 69, 2110.

**Cruz**, L.J., Kupryszewski, G., LeCheminant, G.W., Gray, W.R., Olivera, B.M. and Rivier, J. (1989) "μ-Conotoxin GIIIA, a peptide ligand for muscle sodium channels: chemical synthesis, radiolabeling, and receptor characterization" Biochemistry, 28, 3437-3442.

**Cunningham**, B.C., Lowe, D.G., Li, B., Bennett, B.D. and Wells, J.A. (1994) "Production of an atrial natriuretic peptide variant that is specific for type A receptor" EMBO Journal. 13, 11, 2508-2515.

**Davidson**, A.H., Drummond, A.H., Galloway, W.A. and Whittaker, M. (1997) "The inhibition of matrix metalloproteinase enzymes" Chemistry and Industry 7, 258-261.

**Desmyter**, A., Transue, T.R., Ghahroudi, M.A., Dao Thi, M.H., Poortmans, F., Hamers, R., Muyldermans, S. and Wyns, L. (1996) "Crystal structure of a camel single domain $V_H$ antibody fragment in complex with lysozyme" Nature Structural Biology 3, 803-811.

**Doorbar**, J. and Winter, G. (1994) "Isolation of a peptide antagonist to the thrombin receptor using phage display" J. Mol. Biol. 244, 361-369.

**Etzkorn**, F.A., Guo, T., Lipton, M.A., Goldberg, S.D. and Bartlett, P.A. (1994) "Cyclic hexapeptides and chimeric peptides as mimics of tendamistat" J. Am. Chem. Soc. 116, 10412-10425.

**Fainzilber**, M., Nakamura, T., Gaathon, A., Lodder, J.C., Kits, K.S., Burlingame, A.L. and Zlotkin, E. (1995) "A new cysteine framework in sodium channel blocking conotoxins" Biochemistry 34, 8649-8656.

**Farr-Jones**, S., Miljanich, G.P., Nadasdi, L., Ramachandran, J. and Basus, V.J. (1995) "Solution structure of ω-conotoxin MVIIC, a high affinity ligand of P-type calcium channels, using $^1$H NMR spectroscopy and complete relaxation matrix analysis" J. Mol. Biol. 248, 106-124.

**Fenn**, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M. (1989) "Electrospray ionization for mass spectrometry of large biomolecules" Science 246, 64-71.

**Fisch**, I., Kontermann, R.E., Finnern, R., Hartley, O., Soler-Gonzalez, A.S., Griffiths, A.D. and Winter, G. (1996) "A strategy of exon shuffling for making large peptide repertoires displayed on filamentous bacteriophage" Proc. Natl. Acad. Sci. USA. 93, 7761-7766.

**Gibson**, T. J. (1984) Ph.D thesis, University of Cambridge, U.K.

**Giebel**, L.B., Cass, R.T., Milligan, D.L., Young, D.C., Arze, R. and Johnson, C.R. (1995) "Screening of cyclic peptide phage libraries identifies ligands that bind streptavidin with high affinities" Biochemistry 34, 15430-15435.

**Gomis-Rüth**, F.-X., Maskos, K., Betz, M., Bergner, A., Huber, R., Suzuki, K., Yoshida, N., Nagase, H., Brew, K., Bourenkov, G.P., Bartunik, H. and Bode, W. (1997) "Mechanism of inhibition of the human matrix metalloproteinase stromelysin-1 by TIMP-1" Nature 389, 77-81.

**Gray**, W.R., Olivera, B.M. and Cruz, L.J. (1988) "Peptide toxins from venomous *Conus* snails" Annual Review of Biochemistry 57, 665-700.

**Griffiths**, A.D., Williams, S.C., Hartley, O., Tomlinson, I.M., Waterhouse, P., Crosby, W.L., Kontermann, R.E., Jones, P.T., Low, N.W., Allison, T.J., Prospero, T.D., Hoogenboom, H.R., Nissim, A., Cox, J.P.L., Harrison, J.L., Zaccolo, M., Gherardi, E. and Winter, G. (1994) "Isolation of high affinity human antibodies directly from large synthetic repertoires" EMBO Journal 13, 14, 3245-3260.

**Guddat**, L.W., Martin, J.A., Shan, L., Edmundson, A.B. and Gray, W.R. (1996) "Three-dimensional structure of the α-conotoxin GI at 1.2 Å resolution" Biochemistry 35, 11329-11335.

**Han**, K.-H., Hwang, K.-J., Kim, S.-M., Kim, S.-K., Gray, W.R., Olivera, B.M., Rivier, J. and Shon, K.-J. (1997) "NMR structure determination of a novel conotoxin, [Pro 7, 13] αA-conotoxin PIVA" Biochemistry 36, 1669-1677.

**Harder**, J., Bartels, J., Christophers, E. and Schröder, J.-M. (1997) "A peptide antibiotic from human skin" Nature 387, 861.

**Harrison**, P.M. and Sternberg, M.J.E. (1996) "The disulphide beta-cross – from cystine geometry and clustering to classification of small disulphide-rich protein folds" J. Mol. Biol. 264, 3, 603-623.

**Hill**, J.M., Alewood, P.F. and Craik, D.J. (1996) "Three-dimensional solution structure of μ-conotoxin GIIIB, a specific blocker of skeletal muscle sodium channels" Biochemistry 35, 8824-8835.

**Hill**, J.M., Alewood, P.F. and Craik, D.J. (1997) "Solution structure of the sodium channel antagonist conotoxin GS: a new molecular caliper for probing sodium channel geometry" Structure 5, 571-583.

**Hogan**, J.C. (1996) "Directed combinatorial chemistry" Nature 384, Supplement, 17-19.

**Hopkins**, C., Grilley, M., Miller, C., Shon, K.J., Cruz, L.J., Gray, W.R., Dykert, J., Rivier, J., Yoshikami, D. and Olivera, B.M. (1995) "A new family of *Conus* peptides targeted to the nicotinic acetylcholine-receptor" J. Biol. Chem. 270, 38, 22361-22367.

**Hu**, S., Gehrmann, J., Guddat, L.W., Alewood, P.F., Craik, D.J. and Martin, J.L. (1996) "The 1.1 Å crystal structure of the neuronal acetylcholine receptor antagonist, α-conotoxin PnlA from Conus pennaceus" Structure 4, 417-423.

**Isaacs**, N.W. (1995) "Cystine knots" Current Opinion in Structural Biology 5, 391-395.

**Jones**, S. and Thornton, J.M. (1996) "Principles of protein-protein interactions" Proc. Natl. Acad. Sci. USA 93, 13-20.

**Kauffman**, S.A. (1993) "The Origins of Order" Oxford University Press, UK.

**Kohno**, T., Kim, J.I., Kobayashi, K., Kodera, Y., Maeda, T. and Kazuki, S. (1995) "Three-dimensional structure in solution of the calcium channel blocker ω-conotoxin MVIIA" Biochemistry 34, 10256-10265.

**Koivunen**, E., Wang, B. and Ruoslahti, E. (1995) "Phage libraries displaying cyclic peptides with different ring sizes: ligand specificities of the RGD-directed integrins" Bio/Technology 13, 265-270.

**Kong**, H.M., Kucera, R.B. and Jack, W.E. (1993) "Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis* – Vent DNA polymerase, steady-state kinetics, thermal stability, processivity, strand displacement, and exonuclease activities" J. Biol. Chem. 268, 3, 1965-1975.

**Krätzner**, R. and Kolmar, H. (1997) "The cystine-knot motif as a structural scaffold for the presentation of a constrained peptide library" Protein Engineering 10, SS, 85.

**Ku**, J. and Schultz, P.G. (1995) "Alternate protein frameworks for molecular recognition" Proc. Natl. Acad. Sci. USA. 92, 6552-6556.

**Kuby**, J. (1997) "Immunology" 3rd Edition, W. H. Freeman and Company, USA.

**Ladner**, R.C. (1995) "Constrained peptide libraries as binding entities" Trends in Biotechnology 13, 10, 426-430.

**Lancelin**, J.-M., Kohda, D., Tate, S., Yanagawa, Y., Abe, T., Satake, M., Inagaki, F. (1991) "Tertiary structure of conotoxin GIIIA in aqueous solution" Biochemistry 30, 6908-6916.

**Law**, N.C. and Brose, W. (1997) "Novel analgesics" Chemistry and Industry 8, 306-309.

**Lebrun**, B., RomiLebrun, R., MartinEauclaire, M.F., Yasuda, A., Ishiguro, M., Oyama, Y., Pongs, O. and Nakajima, T. (1997) "A four-disulphide-bridged toxin, with high affinity towards voltage-gated $K^+$ channels, isolated from *Heterometrus spinnifer* (Scorpionidae) venom" Biochemical Journal 328, Pt1, 321-327.

**Lei**, S.-P., Lin, H.-C., Wang, S.-S., Callaway, J. and Wilcox, G. (1987) "Characterization of the *Erwinia carotovora pelB* gene and its product pectate lyase" Journal of Bacteriology 169, 9, 4379-4383.

**Le Nguyen**, D., Heitz, A., Chiche, L., Castro, B., Boigegrain, R.A., Favel, A. and Coletti-Previero, M.A. (1990) "Molecular recognition between serine proteases and new bioactive microproteins with a knotted structure" Biochimie 72, 6-7, 431-435.

**Lin**, S.L. and Nussinov, R. (1995) "A disulphide-reinforced structural scaffold shared by small proteins with diverse functions" Nature Structural Biology 2, 10, 835-837.

**Livnah**, O., Stura, E.A., Johnson, D.L., Middleton, S.A., Mulcahy, L.S., Wrighton, N.C., Dower, W.J., Jolliffe, L.K. and Wilson, I.A. (1996) "Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å" Science 273, 464-471.

**Lowman**, H.B., Bass, S.H., Simpson, N. and Wells, J.A. (1991) "Selecting high-affinity binding proteins by monovalent phage display" Biochemistry 30, 10832-10838.

**Lowman**, H.B. and Wells, J.A. (1993) "Affinity maturation of human growth hormone by monovalent phage display" J. Mol. Biol. 234, 564-578.

**Makrides**, S.C. (1996) "Strategies for achieving high-level expression of genes in *Escherichia coli*" Microbiological Reviews 60, 3, 512-538.

**Markland**, W., Ley, A.C. and Ladner, R.C. (1996a) "Iterative optimization of high-affinity protease inhibitors using phage display. 1. Plasmin" Biochemistry 35, 8045-8057.

**Markland**, W., Ley, A.C. and Ladner, R.C. (1996b) "Iterative optimization of high-affinity protease inhibitors using phage display. 2. Plasma kallikrein and thrombin" Biochemistry 35, 8058-8067.

**Martin**, F., Toniatti, C., Salvatti, A.L., Ciliberto, G., Cortese, R. and Sollazzo, M. (1996) "Coupling protein design and *in vitro* selection strategies: improving specificity and affinity of a designed-protein IL-6 antagonist" J. Mol. Biol. 255, 86-97.

**Martin**, F., Toniatti, C., Salvatti, A.L., Venturini, S., Ciliberto, G., Cortese, R. and Sollazzo, M. (1994) "The affinity-selection of a minibody polypeptide inhibitor of human interleukin-6" EMBO Journal 13, 5303-5309.

**Mattila**, P., Korpela, J., Tenkanen, T. and Pitkanen, K. (1991) "Fidelity of DNA synthesis by the *Thermococcus litoralis* DNA polymerase – an extremely heat-stable enzyme with proofreading activity" Nucleic Acids Research 19, 18, 4967-4973.

**McConnell**, S. and Hoess, R.H. (1995) "Tendamistat as a scaffold for conformationally constrained phage peptide libraries" J. Mol. Biol. 250, 460-470.

**McIntosh**, J.M., Hasson, A., Spira, M.E., Gray, W.R., Li, W.Q., Marsh, M., Hillyard, D.R. and Olivera, B.M. (1995) "A new family of conotoxins that blocks voltage-gated sodium-channels" J. Biol. Chem. 270, 28, 16796-16802.

**McPherson**, M. J., Quirke, P. and Taylor, G. R. (1992) "PCR: a practical approach" I.R.L. Press, Oxford University Press, UK.

**Mitchell**, S.S., Shon, K.J., Foster, M.P., Davis, D.R., Olivera, B.M. and Ireland, C.M. (1998) "Three-dimensional solution structure of conotoxin ψ-PIIIE, an acetylcholine gated ion channel antagonist" Biochemistry 37, 1215-1220.

**Mitchell**, S.S., Shon, K.J., Olivera, B. and Ireland, C.M. (1996) "NMR structures of conotoxins" J. Nat. Toxins. 5, 2, 191-208.

**Moore**, G.J. (1994) "Designing peptide mimetics" Trends in Pharmacological Sciences 15, 4, 124-129.

**Muddiman**, D.C., Wunschel, D.S., Liu, C.L., Pasatolic, L., Fox, K.F., Fox, A., Anderson, G.A. and Smith, R.D. (1996) "Characterisation of PCR products from *Bacilli* using electrospray-ionization FTICR mass-spectrometry" Analytical Chemistry 68, 21, 3705-3712.

**Müller**, W. (1986) "Fractionation of DNA restriction fragments with ion-exchangers for high-performance liquid chromatography" Eur. J. Biochem. 155, 203-212.

**Naito**, Y., Ishikawa, K., Koga, Y., Tsuneyoshi, T. and Terunuma, H. (1995) "Molecular-mass measurement of polymerase chain-reaction products amplified from human blood DNA by electrospray-ionization mass-spectrometry" Rapid Communications in Mass Spectrometry 9, 15, 1484-1486.

**Narasimhan**, L., Singh, J., Humblet, C., Guruprasad, K. and Blundell, T. (1994) "Snail and spider toxins share a similar tertiary structure and cystine motif" Structural Biology 1, 12, 850-852.

**Nishiuchi**, Y. and Sakakibara, S. (1982) "Primary and secondary structure of conotoxin GI, a neurotoxic tridecapeptide from a marine snail" FEBS Letters 148, 2, 260-262.

**Nord**, K., Gunneriusson, E., Ringdahl, J., Ståhl, S., Uhlén, M. and Nygren, P.-Å. (1997) "Binding proteins selected from combinatorial libraries of an α-helical bacterial receptor domain" Nature Biotechnology 15, 8, 772-777.

**Nord**, K., Nilsson, J., Nilsson, B., Uhlén, M. and Nygren, P.-Å. (1995) "A combinatorial library of an α-helical bacterial receptor domain" Protein Engineering 8, 6, 601-608.

**Nordhoff**, E., Cramer, R., Karas, M., Hillenkamp, F., Kirpekar, F., Kristiansen, K. and Roepstorff, P. (1993) "Ion stability of nucleic acids in infrared matrix-assisted laser desorption/ionization mass spectrometry" Nucleic Acids Research 21, 15, 3347-3357.

**Noren**, C.J. (1996) "Phage display: affinity selection linked to genetic information" The NEB Transcript, New England Biolabs Scientific Newsletter, 8, 1, 1-5.

**Nygren**, P.-Å. and Uhlén, M. (1997) "Scaffolds for engineering novel binding sites in proteins" Current Opinion in Structural Biology 7, 463-469.

**Olamendiportugal**, T., Gomezlagunas, F., Gurrola, G.B. and Possani, L.D. (1996) "A novel class of $K^+$-channel blocking toxin from the scorpion *Pandinus imperator*" Biochemical Journal 315, Pt3, 977-981.

**Oliphant**, A.R., Nussbaum, A.L. and Struhl, K. (1986) "Cloning of random-sequence oligodeoxynucleotides" Gene 44, 2-3, 177-183.

**Olivera**, B.M., Hillyard, D.R., Marsh, M. and Yoshikami, D. (1995) "Combinatorial peptide libraries in drug design: lessons from venomous cone snails" Trends in Biotechnology 13, 10, 422-426.

**Olivera**, B.M., Hillyard, D.R., Rivier, J., Woodward, S.R., Gray, W.R., Corpuz, G. and Cruz, L.J. (1990b) "Conotoxins: targeted peptide ligands from snail venoms" American Chemical Society Symposium Series 418, 256-278.

**Olivera**, B.M., Rivier, J., Clark, C., Ramilo, C.A., Corpuz, G.P., Abogadie, F.C., Mena, E., Woodward, S.R., Hillyard, D.R. and Cruz, L.J. (1990a) "Diversity of *Conus* neuropeptides" Science 249, 257-263.

**Olivera**, B.M., Rivier, J., Scott, J.K., Hillyard, D.R. and Cruz, L.J. (1991) "Conotoxins" J. Biol. Chem. 266, 33, 22067-22070.

**Pallaghy**, P.K., Duggan, B.M., Pennington, M.W. and Norton, R.S. (1993) "Three-dimensional structure in solution of the calcium-channel blocker ω-conotoxin" J. Mol. Biol. 234, 405-420.

**Pallaghy**, P.K., Nielsen, K.J., Craik, D.J. and Norton, R.S. (1994) "A common structural motif incorporating a cystine knot and a triple-stranded beta-sheet in toxic and inhibitory polypeptides" Protein Science 3, 10, 1833-1839.

**Palmer**, S.J., Redfern, M.R., Smith, G.C. and Cox, J.P.L. (1998) "Sticky Egyptians: a technique for assembling genes encoding constrained peptides of variable length" Nucleic Acids Research 26, 11, 2560-2564.

**Pörschke**, D. (1977) "Elementary steps of base recognition and helix-coil transitions in nucleic acids" Mol. Biol. Biochem. Biophys. 24, 191-218.

**Pricecarter**, M., Gray, W.R. and Goldenberg, D.P. (1996) "Folding of ω-conotoxins – efficient disulphide-coupled folding of mature sequences *in vitro*" Biochemistry 35, 48, 15537-15546.

**Privalov**, P.L. and Gill, S.J. (1988) "Stability of protein structure and hydrophobic interaction" Advances in Protein Chemistry 39, 191-234.

**Ramilo**, C.A., Zafaralla, G.C., Nadasdi, L., Hammerland, L.G., Yoshikami, D., Gray, W.R., Kristipati, R., Ramachandran, J., Miljanich, G., Olivera, B.M. and Cruz, L.J. (1992) "Novel α-conotoxin and ω-conotoxin from *Conus striatus* venom" Biochemistry 31, 41, 9919-9926.

**Rajewsky**, K. (1996) "Clonal selection and learning in the antibody system" Nature 381, 751-758.

**Rees**, A.R., Staunton, D., Webster, D.M., Searle, S.J., Henry, A.H. and Pedersen, J.T. (1994) "Antibody design: beyond the natural limits" Trends in Biotechnology 12, 199-206.

**Richardson**, J.S. (1981) "The anatomy and taxonomy of protein structure" Advances in Protein Chemistry 34, 167.

**Rigby**, A.C., Baleja, J.D., Furie, B.C. and Furie, B. (1997) "Three-dimensional structure of a γ-carboxyglutamic acid-containing conotoxin, conantokin G, from the marine snail *Conus geographus*: the metal-free conformer" Biochemistry 36, 6906-6914.

**Röttgen**, P. and Collins, J. (1995) "A human pancreatic secretory trypsin inhibitor presenting a hypervariable highly constrained epitope via monovalent phagemid display" Gene 164, 243-250.

**Sambrook**, J., Fritsch, E.F. and Maniatis, T. (1989) "Molecular cloning: a laboratory manual" 2nd Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

**Sanger**, F., Nicklen, S. and Coulson, A.R. (1977) "DNA sequencing with chain-terminating inhibitors" Proc. Natl. Acad. Sci. USA. 74, 12, 5463-5467.

**Sato**, K., Ishida, Y., Wakamatsu, K., Kato, R., Honda, H., Ohizumi, Y., Nakamura, H., Ohya, M., Lancelin, M., Kohda, D. and Inagaki, F. (1991) "Active site of μ-conotoxin GIIIA, a peptide blocker of muscle sodium channels" J. Biol. Chem. 266, 26, 16989-16991.

**Schumacher**, T.N.M., Mayr, L.M., Minor, D.L., Milhollen, M.A., Burgess, M.W. and Kim, P.S. (1996) "Identification of D-peptide ligands through mirror-image phage display" Science 271, 1854-1857.

**Shaler**, T.A., Tan, Y., Wickham, J.N., Wu, K.J. and Becker, C.H. (1995) "Analysis of enzymatic DNA sequencing reactions by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry" Rapid Communications in Mass Spectrometry 9, 942-947.

**Sher**, E., Canal, N., Piccolo, G., Golti, C., Scoppelta, C., Evoli, A. and Clementi, F. (1989) "Specificity of calcium channel auto-antibodies in Lambert-Eaton myasthenic syndrome" Lancet 2, 640-643.

**Shon**, K.J., Stocker, M., Terlau, H., Stühmer, W., Jacobsen, R., Walker, C., Grilley, M., Watkins, M., Hillyard, D.R., Gray, W.R. and Olivera, B.M. (1998) "κ-conotoxin PVIIA is a peptide inhibiting the *Shaker* $K^+$ channel" J. Biol. Chem. 273, 1, 33-38.

**Smith**, C.U.M. (1996) "Elements of molecular neurobiology" 2nd Edition, John Wiley & Sons Ltd., Chichester, England.

**Smith**, G.P. (1994) "The progeny of sexual PCR" Nature 370, 324-325.

**Smith**, G.P. (1985) "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface" Science 228, 1315-1317.

**Smith**, G.P., Patel, S.U., Windass, J.D., Thornton, J.M., Winter, G. and Griffiths, A.D. (1998) "Small binding proteins selected from a combinatorial repertoire" J. Mol. Biol. 277, 2, 317-332.

**Smith**, G.P. and Petrenko, V.A. (1997) "Phage display" Chemical Reviews 97, 391-410.

**Stemmer**, W.P.C. (1994a) "DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution" Proc. Natl. Acad. Sci. USA 91, 10747-10751.

**Stemmer**, W.P.C. (1995a) "The evolution of molecular computation" Science 270, 1510.

**Stemmer**, W.P.C. (1995b) "Searching sequence space" Bio/Technology 13, 549-553.

**Swanson**, G. (1996) "Neurotoxins" 2nd Edition, Trends in Neurosciences, Supplement, 1-37.

**Tindall**, K.R. and Kunkel, T.A. (1988) "Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase" Biochemistry 27, 16, 6008-6013.

**Tonegawa**, S. (1983) "Somatic generation of antibody diversity" Nature 302, 575-581.

**Tramontano**, A., Bianchi, E., Venturini, S., Martin, F., Pessi, A. and Sollazzo, M. (1994) "The making of the minibody: an engineered β-protein for the display of conformationally constrained peptides" J. Mol. Recognit. 7, 9-24.

**Verdine**, G.L. (1996) "The combinatorial chemistry of nature" Nature 384, Supplement, 11-13.

**Vieira**, J. and Messing, J. (1987) "Production of single-stranded plasmid DNA" Methods in Enzymology 153, 3-11.

**Vita**, C., Roumestand, C., Toma, F. and Menez, A. (1995) "Scorpion toxins as natural scaffolds for protein engineering" Proc. Natl. Acad. Sci. USA. 92, 6404-6408.

**Wagner**, S. D., Milstein, C., Neuberger, M. S. (1995) "Codon bias targets mutation" Nature 376, 732.

**Waterhouse**, P., Griffiths, A.D., Johnson, K.S. and Winter, G. (1993) "Combinatorial infection and *in vivo* recombination – a strategy for making large phage antibody repertoires" Nucleic Acids Research 21, 9, 2265-2266.

**Wells**, J.A. (1996) "Hormone mimicry" Science 273, 449-450.

**Winter**, G., Griffiths, A.D., Hawkins, R.E. and Hoogenboom, H.R. (1994) "Making antibodies by phage display technology" Annual Review of Immunology 12, 433-455.

**Woodward**, S.R., Cruz, L.J., Olivera, B.M. and Hillyard, D.R. (1990) "Constant and hypervariable regions in conotoxin propeptides" EMBO Journal. 9, 4, 1015-1020.

**Wrighton**, N.C., Farrell, F.X., Chang, R., Kashyap, A.K., Barbone, F.P., Mulcahy, L.S., Johnson, D.L., Barrett, R.W., Jolliffe, L.K. and Dower, W.J. (1996) "Small peptides as potent mimetics of the protein hormone erythropoietin" Science 273, 458-463.

**Wunschel**, D.S., Muddiman, D.C., Fox, K.F., Fox, A. and Smith, R.D. (1998) "Heterogeneity in *Bacillus cereus* PCR products detected by ESI-FTICR mass spectrometry" Analytical Chemistry 70, 6, 1203-1207.

**Yanisch-Perron**, C., Vieira, J. and Messing, J. (1985) "Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors" Gene 33, 103-119.

# PUBLICATIONS

2560–2564 Nucleic Acids Research, 1998, Vol. 26, No. 11 

# Sticky Egyptians: a technique for assembling genes encoding constrained peptides of variable length

**Simon J. Palmer, Mark R. Redfern, Geoff C. Smith[1] and Jonathan P. L. Cox***

Department of Chemistry and [1]Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

Received March 6, 1998; Accepted April 20, 1998

## ABSTRACT

**Naturally occurring peptides, such as those produced by the poisonous marine snails of the genus *Conus*, have the ability to form tight, highly specific molecular interactions. The rigidity of the peptide framework which promotes these interactions is usually maintained by disulphide bonds, and it seems that the overall main chain conformation (or fold) of the peptide is determined by its length and the sequence distribution of the pairs of cysteine residues participating in these bonds. The fold of the peptide in turn is largely responsible for its shape. Since highly effective molecular interactions occur between species complementary in shape, we reasoned that peptides with the greatest potential in therapy or diagnosis would be found in a library of shapes, those peptides with a shape complementary to a given target being identified, for example, by selection. As a first step towards constructing such a peptide shape library, we have developed a method for assembling DNA fragments which encode an even number of cysteine residues and which are of variable length. We describe this method here.**

## INTRODUCTION

We are interested in exploiting the natural chemical diversity of peptides for medical applications. With this interest in mind, one important consideration is how strongly the peptide interacts with its target. Tight binding requires the peptide to be complementary to the target and to have a rigid architecture. Both factors influence the free energy change accompanying complex formation: complementarity encourages favourable enthalpic contributions (e.g. from van der Waals interactions); rigidity minimizes adverse contributions from conformational entropy losses. Therefore for clinical situations in which tight binding is critical, it would be advisable to begin with a peptide which is complementary to the desired target, and one which has a fixed shape.

The shape of a peptide is broadly dependent on its main chain conformation. This may be constrained and, therefore the shape of the peptide fixed, by one or more disulphide bonds. Structural work on the conotoxins (1), neurotoxic peptides from cone snail venom, and inspection of their amino acid sequences (2), suggest that the particular conformation, or fold, adopted depends upon the length of the sequence and the distribution of the disulphide bonded cysteine residues within that sequence. Given this dependence, which is not unlike that seen in the antigen-binding loops of antibodies (3), we thought that it might be possible to generate a large variety of peptide shapes by altering the position of a small, even number of cysteine residues within short amino acid sequences of varying length. This 'shape library' might then be used to select for peptides complementary to a given target. Once selected, a peptide could be subjected to mutagenesis to improve the strength of its interaction with the target. Such a strategy, again invoking antibodies, would be very similar to the affinity maturation process of the immune system (4).

This paper describes a scheme for assembling DNA fragments encoding a peptide library of the type outlined above (for conciseness we refer to these fragments as genes), the intention being to include these genes in a format suitable for peptide selection, e.g. phage display. The scheme (Fig. 1) involves the ligase-catalysed polymerization of sticky-ended double-stranded oligonucleotide building blocks and hinges upon the fact that there are two codons for cysteine, 5′-UGU-3′ and 5′-UGC-3′.

There are four main types of building block in the scheme, two which propagate polymerization (TGT and TGC, the Egyptians of the title) and two which terminate polymerization (N-CAP and C-CAP). The blocks bear 3 nt overhangs representing one of the two cysteine codons. (Propagating blocks are named according to the sequence of their 5′-overhang, e.g. a TGC block has a 5′-TGC-3′ overhang; see also legend to Fig. 2A for block nomenclature.) These overhangs are arranged such that formation of genes encoding an even number of cysteines is favoured, thereby permitting an integral number of disulphide bonds in the encoded peptide. Each block also possesses a spacer (Fig. 1). The spacers allow the length of the encoded peptide to vary and also most arrangements of cysteines. Spacers for the propagating blocks encode two or more amino acids. [Note that a propagating block with a spacer encoding one amino acid is likely to be unstable (5) and a propagating block 'encoding' 0 amino acids would be non-existent. We did, however, explore the possibility of using a spacer encoding one amino acid in this work.] Spacers for the terminating blocks encode one or more amino acids, or may be absent altogether. Arrangements involving CC and CXC (where X is a single amino acid) are catered for by supplementary propagating blocks (Fig. 1, Box) which insert between two standard blocks while still preserving an even number of cysteines. However, there are some arrangements which the

*To whom correspondence should be addressed. Tel: +44 1225 826548; Fax: +44 1225 826231; Email: j.p.l.cox@bath.ac.uk

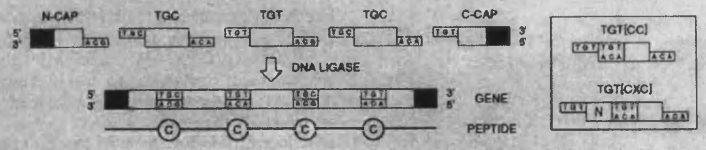



**Figure 1.** Scheme for constructing variable length genes encoding an even number of cysteine residues. Dark rectangle, restriction site for vector incorporation; clear rectangle, spacer; encircled C, cysteine residue. The assembled gene shown encodes four cysteine residues, but genes encoding two, six, eight... cysteine residues may also be formed. Box, supplementary propagating blocks providing for the cysteine arrangements CC and CXC, where X is a single amino acid and N represents 3 bp. Most cysteine arrangements may be generated with these blocks plus four others: equivalent TGC[CC] and TGC[CXC] blocks, a TGT[CC] block in which the internal TGT/ACA section and the spacer have exchanged positions and a TGT[CXC] block in which N and the spacer have exchanged positions. Spacers encode two, three, four... amino acids.

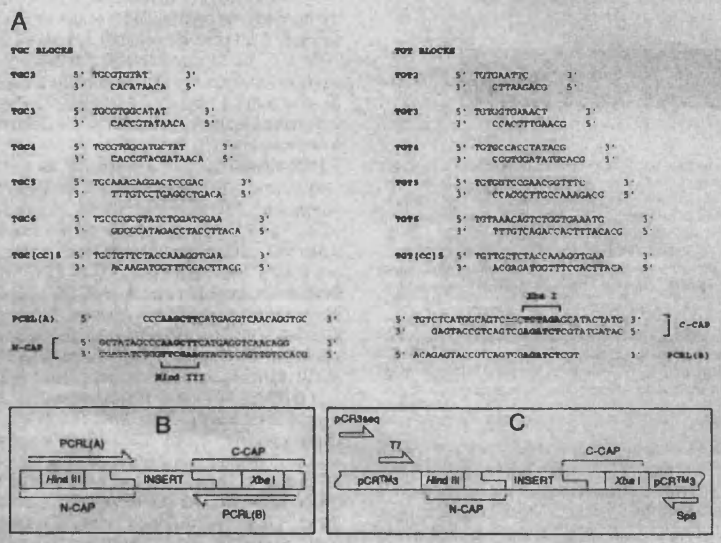



**Figure 2.** (A) Blocks and primers used for this work. Propagating blocks are designated according to the sequence of their 5′-overhang and the number of amino acids encoded by the double-stranded spacer. The supplementary propagating blocks encoding two consecutive cysteines at their 'N-terminus' have an additional [CC] preceding the numeral; similarly, supplementary blocks encoding two cysteines separated by a single amino acid at the 'N-terminus' would be denoted by an additional [CXC] preceding the numeral. In supplementary blocks encoding CC or CXC at their 'C-terminus', the numeral would precede the square brackets. Three other TGC2 blocks were used in which the double-stranded spacer of the original block was replaced by recognition sites for *Xho*I, *Bst*BI and *Eco*RI (5′-CTCGAG-3′, 5′-GAGCGG-3′ and 5′-CTCTTC-3′ respectively). The TGC1 block comprised 5′-TGCTTG-3′ (top strand) and 3′-AACACA-5′ (bottom strand). (B) Assembled gene and primers for its amplification. The last 3 nt of both PCRL(A) and PCRL(B) are complementary to the 3 nt encoding the terminal cysteines of the peptide. (C) Primers for amplification and sequencing of the gene ligated into pCR™3.

scheme does not allow, specifically those involving only two consecutive cysteines (e.g. CC...CC...CC) or two cysteines separated by a single amino acid (e.g. CXC...CXC...CXC) or both (e.g. CC...CXC...CC). The scheme also forbids three or more consecutive cysteines (as may be found in one conotoxin; 2).

Below we present experimental evidence supporting the validity of the proposed scheme.

## MATERIALS AND METHODS

Standard recombinant DNA techniques (6) were used throughout unless otherwise stated. Restriction enzymes, T4 polynucleotide kinase and T4 DNA ligase were purchased from New England Biolabs. Taq DNA polymerase, used for all PCRs, was obtained from Bioline. Oligonucleotides were purchased from Perkin-Elmer

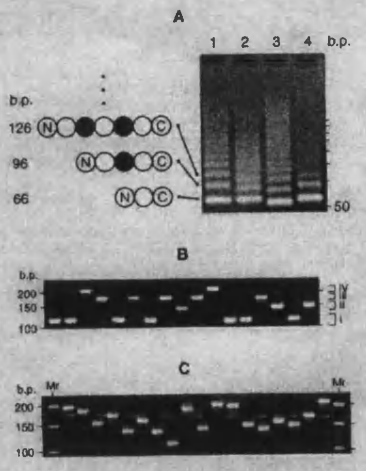



**Figure 3.** Experimental verification of the proposed scheme. (**A**) PCR products of various ligation mixtures. Lane 1, TGC4 and TGT4. Each band corresponds to a gene encoding an even number of cysteine residues. The three smallest genes are indicated schematically on the left side of the figure (encircled N, N-CAP; encircled C, C-CAP; open circle, TGC4; filled circle, TGT4). Lane 2, TGC4, TGT4, TGC[CC]5 and TGT[CC]5. Lane 3, TGC2 and TGT4. Lane 4, TGC4 and TGT2. The unmarked gradations to the right side of the photograph correspond to 50 bp increments. (**B**) PCR screen of 17 transformants arising from a restriction digested TGC4/TGT4 ligation mixture cloned into the vector pCR™3. Brackets I–IV identify bands corresponding to genes encoding two, four, six and eight cysteines respectively (constructs I–IV in the text). (**C**) PCR screen of 18 transformants arising from a restriction digested TGC2–6/TGT2–6/TGC[CC]5/TGT[CC]5 ligation mixture cloned into pCR™3. Mr, 50 bp DNA ladder (Gibco). All PCR products were run on 3% agarose gels.

Applied Biosystems. DNA sequencing was performed on an ABI PRISM 377 DNA Sequencer.

Double-stranded oligonucleotide building blocks for the ligation reactions were constructed by annealing the appropriate oligonucleotides. All oligonucleotides were phosphorylated at the 5′-terminus, apart from the upper strand of N-CAP and the lower strand of C-CAP. Oligonucleotides for the TGC4/TGT4 blocks (see Fig. 2A and its legend for a description of the blocks used in this work) employed in the ligation leading to the PCR screen shown in Figure 3B were purchased pre-phoshorylated. The remaining oligonucleotides were phosphorylated with T4 polynucleotide kinase according to the supplied instructions, except that, in preparation for ligation, T4 DNA ligase buffer was used. Annealing was performed by mixing an equal number of moles of each of the oligonucleotides constituting a particular block, making the solution up to 25 µl with T4 DNA ligase buffer and water and then lyophilizing. When mixing the oligonucleotides, care was taken to keep the volume of residual glycerol (from the T4 polynucleotide kinase storage solution) below 0.05 µl. Residual volumes greater than this prevented proper lyophilization. Each annealed block was dissolved in 25 µl water at 4°C.

Most ligations were performed with a ratio of N-CAP:TGC:TGT:C-CAP of essentially 1:10:6:1. For the ligations leading to the PCR products shown in lanes 1, 3 and 4 of Figure 3A, concentrations were (N-CAP:TGC:TGT:C-CAP) 1:10:6:1 pmol. For the ligation leading to the PCR products shown in lane 2 of Figure 3A, concentrations were (N-CAP:TGC4:TGT4:TGC[CC]5:TGT[CC]5:C-CAP) 1:6:2:2:2:1 pmol, maintaining the concentrations of propagating block overhangs implicit in the original 1:10:6:1 ratio. For the ligation leading to the PCR screen of Figure 3B, the concentrations of (N-CAP:TGC:TGT:C-CAP) were increased to 5.5:55:33:5.5 pmol, so that the mixture could be visualized on a gel. For the initial multi-block ligations involving TGC2–6, TGT2–6, TGC[CC]5 and TGT[CC]5, the concentration of each standard TGC block was 10 pmol and of each TGT, TGC[CC]5 and TGT[CC]5 block 5 pmol, with N-CAP and C-CAP 6 pmol. In subsequent multi-block ligations involving the TGC2 blocks containing a recognition site for a restriction enzyme, the concentration of TGC2 was 30 pmol, TGC3–6 each 6 pmol, TGT2 18 pmol, TGT3–6, TGC[CC]5 and TGT[CC]5 each 3.6 pmol and N-CAP and C-CAP both 6 pmol. To perform the ligations, the propagating blocks were mixed, T4 DNA ligase added (4 U/µl reaction volume) and the mixture left for 1 h at 0 (TGC1/TGT4 ligation) or 4°C (all other ligations). (Ligations were performed at low temperature to enhance the stability of the blocks.) Subsequently N-CAP and C-CAP were added, together with more T4 DNA ligase (200 U), and the new mixture left for a further 2 h at 4°C.

PCR amplification of ligation mixtures (0.5 µl) was achieved with primers PCRL(A) and PCRL(B) (Fig. 2B) using the programme: [94°C, 1 min; 65°C, 1 min; 72°C, 1 min] × 25; 72°C, 5 min. Including the two 3 bp terminal joints, amplification with these primers appended 54 bp to the insert between N-CAP and C-CAP [e.g. for a TGC4:TGT4:TGC4 insert, the size of the PCR product would be 54 + (3 × 12) + (2 × 3) = 96 bp]. After alcohol precipitation, the relevant ligations were restriction digested with *Hin*dIII and *Xba*I and purified by agarose gel electrophoresis, excising DNA in the range 50–150 bp (TGC4/TGT4 ligation) and 60–150 bp (TGC2–6/TGT2–6/TGC[CC]-5/TGT[CC]5 ligations). Following digestion, the mixture was ligated into *Hin*dIII- and *Xba*I-cut pCR™3 (Invitrogen) and transformed into *Escherichia coli*. PCR screening of ampicillin-resistant colonies was performed with the primers T7 and Sp6 (T7, 5′-TAATACGACTCACTATAGGG-3′; Sp6, 5′-GATTTAGGTGACACTATAG-3′) (Fig. 2C), using the programme: 94°C, 5 min; [94°C, 1 min; 55°C, 1 min; 72°C, 1 min] × 30; 72°C, 5 min. Together with the two terminal 3 bp joints, these primers appended 103 bp to the insert. Restriction digestions of inserts amplified in PCR screens were carried out after they were alcohol precipitated. Sequencing was performed with the primer pCR3seq, 5′-AGGTCTATATAAGCAGAGCT-3′ (Fig. 2C).

## RESULTS

To test the principle of our scheme, we performed a ligation with TGT and TGC blocks comprising 12 putative base pairs (TGC4 and TGT4 respectively, where 4 represents the number of amino acids encoded by the spacer; see legend to Fig. 2A). Amplification of a small part of the ligation mixture with primers specific for N-CAP and C-CAP gave a ladder of bands (Fig. 3A, lane 1). Each band corresponded in size to a gene encoding an even number of cysteine residues (Fig. 3A, schematic). A similar TGC4/TGT4 ligation mixture was restriction digested, purified on an agarose gel, isolating DNA expected to include genes encoding two, four, six and eight cysteines only (designated constructs I–IV respectively), and inserted into a vector. The vector population was transformed

into *E.coli* and transformants screened for inserts (Fig. 3B). Of 35 screened transformants, 33 gave bands corresponding to genes encoding an even number of cysteine residues, with a bias towards construct I (I, 14; II, 6; III, 8: IV, 5). Two bands were larger than expected and probably corresponded to two constructs ligated together. We did not encounter this problem again. Twelve (three of each construct) inserts were sequenced: each sequence confirmed the fidelity of the assembly process.

Having shown successful assembly for two particular propagating blocks, we then went on to investigate the stability, under the ligation conditions used, of two of the smallest double-stranded propagating blocks (TGC2 and TGT2) needed for the scheme. We did two ligations, one with TGC2 and TGT4 and one with TGT2 and TGC4. On amplification, each ligation mixture gave a ladder diagnostic of correct gene formation (Fig. 3A, lanes 3 and 4), although the extent of large construct formation was greater for TGC2/TGT4 (lane 3). Correct ladder formation was not observed when a TGC1/TGT4 ligation mixture was PCR amplified (data not shown).

Subsequently we investigated the ability of the supplementary blocks to insert between standard blocks by performing a TGC4/TGT4 ligation in the presence of TGC[CC]5 and TGT[CC]5. Amplification of this mixture gave a ladder consistent with correct insertion events (Fig. 3A, lane 2), with the lowest band unaffected and the higher bands smeared. (The second lowest band probably corresponds to constructs containing a standard propagating block ligated to a supplementary one.)

Finally, we performed a ligation involving TGC2–6, TGT2–6, TGC[CC]5 and TGT[CC]5. Following restriction digestion of the ligation mixture, DNA expected to include genes encoding between four and twelve cysteine residues was isolated. We deliberately tried to avoid genes encoding two cysteines in the light of the bias towards construct I seen in the TGC4/TGT4 experiment described above. (These genes could in any case be obtained in a separate ligation involving the TGC2–6 blocks alone.) Despite this precaution, about a quarter of the bands from PCR screens were suggestive of genes encoding two cysteines. The remaining bands were within the expected upper size limit; the bands were also encouragingly varied in size (Fig. 3C).

Fifty nine inserts from eight independent ligations were sequenced. The majority of these sequences (fifty four) showed that polymerization had occurred as desired; the remaining five sequences indicated undesirable assembly events (e.g. TGC6 ligated to TGC5). Forty three of the fifty four correctly assembled genes were unique. Concatamers of the supplementary blocks were observed in only two cases, both involving TGT[CC]5 and comprising two and three blocks (we wanted to avoid extensive concatenation as it would give genes encoding dense arrays of cysteines, leading to peptides unlikely to fold productively). Inserts which had assembled correctly contained one, two, three, four, five, six and seven propagating blocks (five, ten, twenty, six, seven, five and one inserts respectively) and encoded two, four, six, eight and ten cysteines (five, twenty nine, thirteen, six and one insert respectively). Theoretically we might also have encountered inserts with 8–10 propagating blocks.

The sequences also revealed certain trends for the individual blocks. Thus TGT5 was rarely incorporated, while TGC5, although it occurred regularly, was not often found ligated to N-CAP. TGC[CC]5, on the other hand, was *only* found ligated to N-CAP, unlike its sister block TGT[CC]5, which did insert between the standard propagating blocks and was also found ligated to C-CAP. TGC[CC]5 was able to self-polymerize in the absence of other blocks, however (data not shown).

Far more worrying was the complete absence of TGC2 in inserts from the initial multi-block ligations and the infrequent appearance of TGT2 (about one in twelve inserts). In an attempt to rectify this situation, the concentration of both blocks was raised to five times that of their larger counterparts. This increased the frequency with which TGT2 occurred to a more acceptable level (approximately one in seven inserts), but TGC2 was still undetectable. To enable large numbers of inserts to be screened for the presence of TGC2, a new TGC2 block was designed, TGC2[Xho] (see legend to Fig. 2A), in which the putative double-stranded region comprised a *Xho*I recognition site (TGT2 already had an *Eco*RI site for this purpose). Like TGT2, this new block had a G:C base pair at either end which we hoped would make the block less susceptible to possible fraying. It also had two internal G:C base pairs, lending it extra stability. Digestion of amplified inserts from multi-block ligations involving TGC2[Xho] did indeed suggest that it had been included in some cases. On sequencing these inserts, however, it was apparent that the oligonucleotides comprising TGC2[Xho] and TGT2 had self-annealed (TGC2[Xho] had also annealed as required) and that these aberrant blocks had facilitated each other's incorporation into the assembly product, severely disrupting the ligation scheme. This mutually assisted misincorporation had not been possible with the previous TGC2/TGT2 pairing since only the TGT2 oligonucleotides could self-anneal. Self-annealing of the TGT2 oligonucleotides would lower the concentration of the TGT2 block proper, possibly explaining its infrequent inclusion in the multi-block ligations and the attenuated ladder of bands from the PCR-amplified TGC4/TGT2 ligation mixture (Fig. 3A, lane 4).

To resolve the self-annealing problem in the TGC2 block (we were less concerned about TGT2 since we had already observed its correct incorporation) two new blocks were devised in which the palindromic *Xho*I site was replaced with the asymmetrical recognition sequences of *Bsr*BI and *Ear*I (see legend to Fig. 2A). Like TGC2[Xho], both blocks have terminal and internal G:C pairs which ought to make them relatively robust. In restriction digestion screens of amplified inserts from multi-block ligations involving either TGC2[Bsr] or TGC2[Ear], both sites were detected at satisfactory frequencies (about one in five inserts and one in three inserts respectively). Sequencing of inserts containing these sites showed that both blocks had incorporated as intended.

## DISCUSSION

We have conceived and tested a scheme for assembling genes encoding an even number of cysteine residues and which vary in length. The scheme involves the ligation of double-stranded blocks of DNA with sticky ends. It works reasonably well. Most assembled genes do encode an even number of cysteines, although some (<10%) exhibit an inappropriate assembly event resulting in these genes encoding an odd number of cysteines. The genes encode quite diverse arrangements of cysteines and their length, which may be controlled at the stage when the assembled DNA fragments are isolated by gel electrophoresis, also shows good variation (Fig. 3C).

Nearly all the blocks with which we attempted the scheme participated in gene assembly. These blocks had spacers which encoded between two and six amino acids. (A block whose spacer encoded one amino acid, TGC1, did not participate in assembly,

as expected.) The behaviour of two of the smallest blocks, TGC2 and TGT2, and a supplementary block, TGC[CC]5, was of some concern, however. In order to detect TGC2 and TGT2 at frequencies comparable with the other blocks, the concentrations of both had to be raised relative to these blocks. In addition, only those TGC2 blocks with a G:C base pair at either end of the double-stranded region were observed (the TGT2 block fortuitously already had G:C base pairs at these locations). If one wanted to include some sequence variation in the smallest blocks, this apparent G:C requirement would restrict the first amino acid encoded by the blocks to just ten (L, P, H, Q, R, V, A, D, E and G), assuming those blocks were made by annealing complementary oligonucleotides, as they were here (see Materials and Methods), although it would not limit the identity of the second amino acid. However, the sequence diversity encoded by the genes is more critically affected by the format in which their peptide products are selected (see final paragraph).

The other problematical block, TGC[CC]5, was only found ligated to N-CAP and could not be detected inserted between two standard propagating blocks. Again, although irksome, this preference would not be fatal to the scheme, which may be realized with three alternative sets of supplementary blocks, (TGT-[CC][$X_{1-6}$], TGT[CXC][$X_{1-6}$], TGT[$X_{1-6}$][CC], TGT[$X_{1-6}$]-[CXC], TGC[CC][$X_{1-6}$] and TGC[CXC][$X_{1-6}$]) or (TGC-[CC][$X_{1-6}$], TGC[CXC][$X_{1-6}$], TGC[$X_{1-6}$][CC], TGC[$X_{1-6}$]-[CXC], TGT[$X_{1-6}$][CC] and TGT[$X_{1-6}$][CXC]) or a combination of all eight types of supplementary block (see legend to Fig. 2A for supplementary block nomenclature). Using the first set of supplementary blocks, the scheme requires that the TGC supplementary blocks ligate to N-CAP alone or as self-polymerized units (and TGC[CC]5 does self-polymerize) and that the TGT supplementary blocks ligate to C-CAP as well as insert between the standard propagating blocks, which TGT[CC]5 can do.

Overall, therefore, it seems that the scheme could be used to construct genes encoding most combinations of cysteines. If the peptides in a potential library were restricted in length to ≤30 residues, with 0–6 residues between consecutive cysteines, roughly in keeping with the conotoxins (2), the maximum number of folds attainable with the scheme is ~$10^7$. [The number of folds = number of possible cysteine arrangements × number of ways of pairing cysteines, by disulphide bonds, in a particular arrangement. The number of possible cysteine arrangements in a sequence of length $L$ containing $N$ cysteines is the coefficient of $X^{L-N}$ in the *escargot* polynomial $(1 + X + X^2 + ... + X^6)^{N-1}$ once it has been expanded. So for a peptide containing 30 amino acids, eight of which are cysteines, the relevant term is $59\,710X^{22}$ and thus the number of possible cysteine arrangements in this peptide is 59 710. The polynomial arises by considering sequences as products, e.g. $CXXCXCXC = C^4X^4$. In formulating it, it was assumed that amino acids beyond the two terminal cysteines contribute little to the peptide fold; $L$ is therefore defined by these two termini. The polynomial may also be used to compute the number of sequences in the library. The number of ways of pairing $N$ cysteines is given by the expression $(N-1)\times(N-3)\times...\times 1$.] Some of these potential folds will be excluded on stereochemical grounds, while others will be disfavoured energetically. Therefore the actual number of folds will almost certainly be $<10^7$.

Sequence variation could be introduced into the library with blocks encoding random amino acids. Theoretically these blocks could give rise to ~$10^{33}$ different sequences in the 30 amino acid size range, even taking into account sequence restrictions imposed on the TGC2 and TGT2 blocks. This greatly exceeds the library size manageable with a standard selection format such as phage display (the size of conventional phage display libraries is ≤$10^8$; 7). Thus the number of amino acid sequences in a library created through this scheme would be only a tiny fraction of those possible. However, the scheme seeks primarily to generate a collection of shapes into which sequence diversity may be subsequently introduced following selection of a shape complementary to a given target. There are clearly limitations to this approach (for example, as a consequence of the poor sequence diversity some shapes may be electrostatically incompatible with the target due to charge repulsions and, therefore may not be selected), but despite these limitations we feel that the approach is one worth pursuing.

## ACKNOWLEDGEMENTS

This work was suggested by a talk given by Bruce Livett at the University of Bath in 1995 and profited from very helpful conversations with Mark Searle, Dek Woolfson and Alan Cox. Financial support from The Royal Society, The Wellcome Trust, the BBSRC and the University of Bath is gratefully acknowledged, as is the use of laboratory space in the Department of Biology and Biochemistry at Bath. Further details of the mathematics used in this paper may be found on the Internet at http://www.bath.ac.uk/~masgcs/escargot.html

## REFERENCES

1 Mitchell,S.S., Shon,K.J., Olivera,B. and Ireland,C.M. (1996) *J. Natural Toxins*, **5**, 191–208.
2 Myers,R.A., Cruz,L.J., RivierJ.E. and Olivera,B.M. (1993) *Chem. Rev.*, **93**, 1923–1936.
3 Chothia,C. and Lesk,A.M. (1987) *J. Mol. Biol.*, **196**, 901–917.
4 Siskind,G.W. and Benacerraf,B. (1969) *Adv Immunol.*, **10**, 1–50.
5 Pörschke,D. (1977) *Mol. Biol. Biochem. Biophys.*, **24**, 191–218.
6 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
7 Waterhouse,P., Griffiths,A.D., Johnson,K.S. and Winter,G. (1993) *Nucleic Acids Res.*, **21**, 2265–2266.